AD_____

Award Number: W81XWH-08-1-0669

TITLE: Discovery of Genomic Breakpoints Affecting Breast Cancer Progression and Prognosis

PRINCIPAL INVESTIGATOR: Petra den Hollander, Ph.D.

CONTRACTING ORGANIZATION: Baylor College of Medicine
Houston, Texas 77030

REPORT DATE: October 2010

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
October 2010	Annual Summary	15 September 2008 – 5 January 2011

4. TITLE AND SUBTITLE

Discovery of Genomic Breakpoints Affecting Breast Cancer Progression and Prognosis

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-08-1-0669

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Petra den Hollander, Ph.D.

E-Mail: pdhollander@mdanderson.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Baylor College of Medicine
Houston, Texas 77030

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

157 genomic breakpoints could be confirmed as likely somatic mutations. We focused on breakpoints predicted to lead to fusion transcripts. By RT-PCR we determined that 4 showed a fusion mRNA. In the case of the ARFGEF2/SULF2, a non-functional Sulfatase 2 might be created. To give insight, SULF2 mRNA was knocked down using siRNA. Cells treated with SULF2 siRNA, exhibited a growth advantage enhanced survival, and an advantage in anchorage-independent growth compared to control siRNA. This implies that the presence of this fusion might mean a loss of function of the tumor suppressor Sulfatase 2. Another fusion, RAD51C/ATXN7 results in the truncation RAD51C, involved in double stranded break repair. We detected chimeric mRNA expression in 3 breast cancer cell lines, and a shorter form of RAD51C by western blot, indicating that the fusion introduces a truncation in RAD51C protein. To gain insight into the heterogeneity of genomic breakpoints, in seven MCF-7 sub-lines. There is an enrichment for breakpoints containing genes (50.3% vs 77.4%), and for fusion-containing breakpoints (6.4% vs 16.1%). When studying cell lines originating from a single cell, we discovered that there is very little genetic variability between them. A large effort has gone into the development of next-generation sequencing techniques for the discovery of genomic breaks and fusion in the breast cancer genome. We developed new techniques and validated them with standard PCR. The validation rate is very promising, and these new techniques will aid us in the discovery of breast cancer.

15. SUBJECT TERMS
Breast Cancer, Genomic Breakpoints, Tumor suppressor, evolution next-generation sequencing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	34	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Introduction

Over the past decade, genetic changes associated with recurrent chromosome breakpoints have been discovered in human malignancies, predominantly of haematologic origin. The characterization of these alterations has demonstrated that these changes can be disease specific and can have functional consequences (e.g. Philadelphia chromosome and CML). ***The rationale underlying this proposal is that functional recurrent breakpoint-associated chromosomal changes occur during breast cancer progression and that their discovery and characterization may lead to novel diagnostic and therapeutic tools for breast cancer patients.***
By combining studies in cell lines, and breast tumors, we can identify important genomic rearrangements that may drive breast cancer tumorigenesis. With these studies we push the envelope in developing new screening methods for more efficient, informative and cost-effective discovery of recurrent genomic rearrangements in breast cancer. By using unbiased approaches to pursue the discovery of important aberrant breakpoints in the breast cancer genome, we widely increase the chance of detecting a rare, but important alteration, and recurrences in different breast cancer genomes. By studying the biology of such an aberrant breakpoint in the genome, it might be feasible to discover new ways of targeting these specific alterations on the protein level with new compounds. Taken together, this study will uniquely determine for the first time recurrent biological relevant genomic changes in breast cancers.

Body

Task 1: *Determine the recurrence rate of the breakpoints in breast cancer cell lines.*

To determine the recurrence rate of the genomic breakpoints in breast cancer cell lines, I created 2 pools with breast cancer cell line DNA. One pool contained 9 cell lines and the other 7. In total 16 cell lines were tested for the presence of the 398 breakpoints. Bands were present in about 30% of the PCR products. These PCRs were performed before we analyzed them in depth in MCF-7. When sequence analyzing the breakpoints in MCF-7/BAC, we discovered that many breaks were induced during the creation of the BAC library. Other breakpoints were eliminated by inability to validate on BAC or MCF-7 cell line DNA, presence in the normal population, or redundancy.
We discovered 157 genomic breakpoints in MCF-7 cells that could be confirmed by PCR across breakpoint joins as likely somatic mutations, and of which only some have been previously described. A total of 79 genes are involved in rearrangement events, including 10 fusions of coding exons from different genes and 77 other aberrant breakpoints involving known or predicted genes. Among the breakpoints that involved genes, we first focused on those 10 gene fusion predicted to lead to fusion transcripts (see Table 1).

Fusion Genes in MCF-7 Cells

ARFGEF2 : *SULF2*	Intra-Chr Inversion	20q13.13;20q13.13	Fusion of *ARFGEF2* exon 1 to *SULF2* exons 3-21, 1.2Mb inversion
DEPDC1B : *ELOVL7*	Intra-Chr Inversion	5q12.1;5q12.1	Fusion of *DEPDC1B* exons 1-7 (out of 11) with *ELOVL7* exons 8-9, 127Kb inversion
RAD51C : *ATXN7*	Inter-Chr Rearrangement	3p14.1;17q22	Fusion of *RAD51C* N-terminus exons 1-7 (out of 9) with *ATXN7* exons 6-13
SULF2 : *PRICKLE2*	Inter-Chr Rearrangement	3p14.1;20q13.13	Fusion of *SULF2* exon 1 with last exon of *PRICKLE2*
NPEPPS : *USP32*	Intra-Chr Inversion	17q21.32;17q23.2	Fusion of *NPEPPS* exons 1-12 (out of 23) with *USP32* exons 2-34, 13Mb inversion
ASTN2 : *PTPRG*	Inter-Chr Rearrangement	3p14.2;9q33.1	Fusion of *ASTN2* exons 1-10 (out of 22) with *PTPRG* exons 3-30
BCAS3 : *BCAS4*	Inter-Chr Rearrangement	17q23.2;20q13.13	*BCAS4* exon 1 fused to *BCAS3* exons 23-24, Ruan *et al.* (Genome Res 17:828-838)
BCAS3 : *RSBN1*	Inter-Chr Rearrangement	1p13.2;17q23.2	Fusion of *RSBN1* first exon with *BCAS3* exons 6-24
ASTN2 : *TBC1D16*	Inter-Chr Rearrangement	9q33.1;17q25.3	Fusion of *ASTN2* exons 1-15 with *TBC1D16* exons 2-12
BCAS4 : *PRKCBP1*	Intra-Chr Inversion	20q13.12;20q13.13	Fusion of *BCAS4* exon 1 with *PRKCBP1* exons 5-22, 3.5Mb inversion

Table1: Gene fusions discovered in MCF-7 breast cancer cell line

For a gene fusion to have a function significance it needs to make an aberrant protein that will have an alternate function then the wildtype proteins. The first step in identifying these is to determine if these gene fusions produce a chimeric mRNA. To determine if the predicted chimeric mRNA transcript was created by these genomic fusions, I performed gene-specific RT-PCR on MCF-7 and 2 normal controls. Out of ten DNA fusions, four showed a fusion mRNA transcript in MCF-7 specifically by RT-PCR (Figure 1).

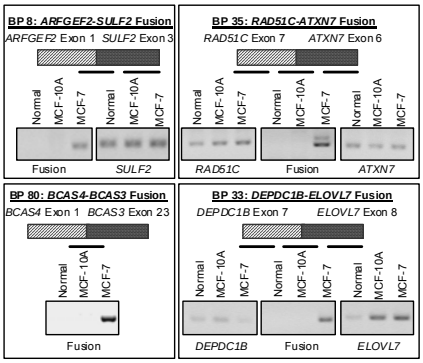

Figure 1: Discovery of chimeric mRNA product in MCF-7. RNA was isolated from MCF-7 cells, and RT-PCR was preformed for control regions and the fusion. These data clearly show the presence of the wildtype transcript in MCF-7 and the two control RNAs (from MCF-10A, and normal breast), while the fusion transcript is only present in MCF-7.

Three of these are newly identified (ARFGEF2/SULF2, DEPDC1B/ELOVL7, RAD51C/ATXN7), and one has been previously described (BCAS4/BCAS3). If a genomic fusion is present in other breast cancer cell lines it is not very likely it will occur exactly at the same position. Even if the break occurs several kilobases up or down stream of the originally discovered breakpoint in MCF-7, it might still create the same down stream consequence by making the same chimeric mRNA. Because of this, I tested the presence of the 4 different chimeric mRNA in 16 breast cancer cell lines. The 16 breast cancer cell lines were divided into 4 pools of 4 cell lines. Pool 1 and 2 showed the presence of a band after performing RT-PCR for the RAD51C/ATXN7 fusion (Figure 2). Sequencing of the PCR product confirmed the presence of the fusion. After deconvoluting the pools, I discovered that the RAD51C/ATXN7 fusion was present in 2 other breast cancer cell lines, T47D, and MDA MB361. The fusion of RAD51C and ATXN7 most likely results in loss of a critical C-terminal domain of RAD51C. By western blot I was able to detect a shorter band in MCF-7 and MDA MB361. This was confirmed by performing an immunoprecipitating RAD51C with a specific antibody, and probing the western blot with a different RAD51C antibody (Figure 3).

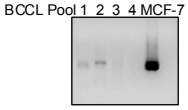

Figure 2: Discovery of chimeric mRNA product of the RAD51C/ATXN7 fusion in breast cancer cell lines. RNA was isolated from 16 breast cancer cell lines, and 4 pools of 4 cell lines were made. RT-PCR was performed the fusion. The presence of the fusion transcript was confirmed by sequencing.

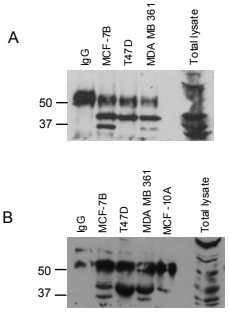

Figure 3: Presence of a truncated form of RAD51C in breast cancer cell lines. A) Immunoprecipitation was performed on cell lysates with a mouse anti-Rad51C antibody. Elutes from the IP were run on an SDS-page and probed with an rabbit anti-Rad51C antibody. B) Repeat of experiment in A, with the addition of a negative control MCF10-A

I performed preliminary functional studies for the ARFGEF2/SULF2 fusion, which are reported under Task 4.

Another angle we pursued is the evolution of breakpoints. To gain insight into the heterogeneity of genomic breakpoints in breast cancer cell lines, and to also narrow down on breakpoints originating from the ancestor, I studied these 157 validated breakpoints in seven MCF-7 sub-lines (Figure 4). Thirty-one breakpoints from the original 157 identified in MCF-7/BAC, are present in all MCF-7 sub-lines (Figure 5). When looking at the distribution of these breakpoints in the genome, we can see that there are clusters of breakpoints on certain chromosomes, and

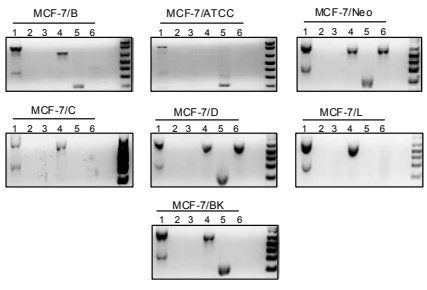

Figure 4: Examples of breakpoint analysis by PCR. Breakpoints were scored on the presence, size and number of PCR bands.

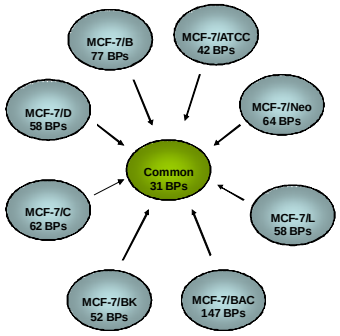

Figure 5: Schema of breakpoints present in individual cell lines and the shared breakpoints between them

that there are many breakpoints randomly distributed throughout the genome. When focusing on the 31 breakpoints that are common in all the MCF-7 sub-lines, it is clear that the clusters are retained and that the amount of breakpoint randomly distributed is dramatically reduced. A finding of interest is that there is a great enrichment of breakpoints containing genes (50.3% vs 77.4%, p=0.0056) (Figure 6). Even more interesting is that 5 of the 10 fusions are in all cell lines (6.4% vs 16.1%) (Figure 6). Also, all 4 fusion genes expressing chimeric mRNA product are present in all MCF-7 sub-lines, indicating that these breakpoints might originate from an ancestral cell line.

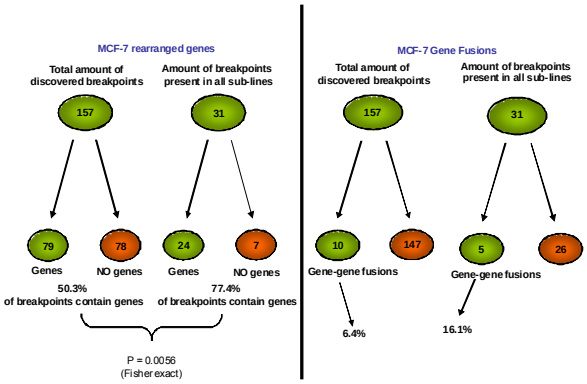

Figure 6: Graphic representation of the distribution of the break points. The enrichment of gene-containing breakpoints is statistically significant (p=0.0056, Fisher exact)

With this information we may get a better understanding of the evolution and heterogeneity of genomic instability and rearrangements in breast cancer. Also, by narrowing down on breakpoints that are in all the MCF-7 sub-lines, I get closer to the 'true' breakpoints that originated in the tumor of which MCF-7 is derived.

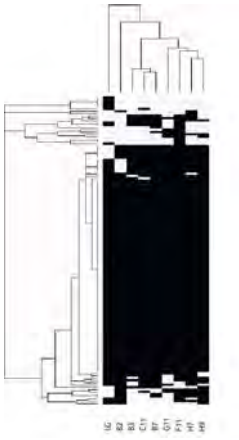

Figure 7: heatmap showing that the clones are closely related to the parental cell line MCF-7LG. Black is present, white is absent breakpoint in that particular cell line.

To address the clonal variation within cell lines, I tested for the presence of the 157 breakpoints in a set of MCF-7 cell lines that were derived from the same parental MCF-7 line MCF-7LG. As you can see in figure 7, there is close to 95% homology between the clones. This is an important finding, since this shows that single cells from a population are much closer related then we expected, and that their genomes are more stable then speculated in the past.

Task 2: *Determine recurrence rate of the breakpoints in a panel of breast tumors.*

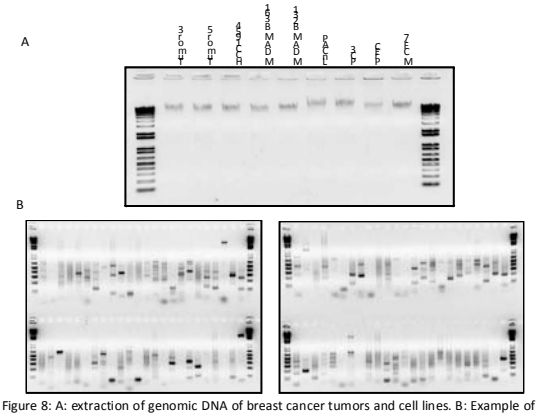

Figure 8: A: extraction of genomic DNA of breast cancer tumors and cell lines. B: Example of validation of breakpoints by PCR in HCC1194

To use genomic DNA for the development of next-generation sequencing techniques the quality of the DNA extracted is crucial. I compared extraction protocols to determine the best technique to get the DNA required for the long range PCR. This DNA was used for de development of new high-throughput sequencing for the discovery of breakpoints in breast cancer. In this case we used SOLiD sequencing. Our intension was to generate a 5 kb mate pair library to increase the sensitivity of the discovery of

breakpoints in the genome. We encountered a lot of difficulty to generate a library with this size insert, and spend a lot of time and resources in the optimization of the library to get an appropriate coverage of the whole genome. We were able to generate libraries from DNA extracted from breast cancer tissue, and its paired normal control. This was extremely valuable, since this gave us insight in the background presence of breakpoints and fusion in the DNA of that individual. We also included three breast cancer cell lines (HCC1954, MDA MB231, and MDA MB361. The data from our SOLiD sequencing looks very promising. We were able to find recurrences of breakpoints in the breast tumors as well as in the breast cancer cell lines. I validated by PCR the identified breakpoints, and we had an unexpected high validation rate of almost 50% (42.3-44.8%).

Task 3: *Validate breakpoints in an independent set of breast cancer tumors and associate breakpoints with histo-pathological and clinical characteristics.*

a. Develop break-away FISH probes for the detection of recurrent breakpoints.

FISH probes were developed for the RAD51C/ATXN7 fusion. After confirming probe was specificity on control lymphocytes, metaphase spreads of MCF-7 and MCF-10A (negative control) cells, were hybridized with probes proximal, and distal of the break in RAD51C, and a probe for RAD51C spanning the break (Figure 9). These break-away FISH experiments did not give a conclusive answer, thus we decided to test for the presence of the fusion. Probes for RAD51C and ATXN7 were developed, and hybridized on MCF-7 and MCF-10A metaphase spreads. These results clearly show the presence of RAD51C and ATXN7 signal in close proximity in the MCF-7 cells and not in the MCF-10A cells (Figure 10). With these data we were able to generate a detection tool for the presence of the RAD51C/ATXN7 fusion, and to confirm the presence of the genomic translocation in MCF-7 cells.

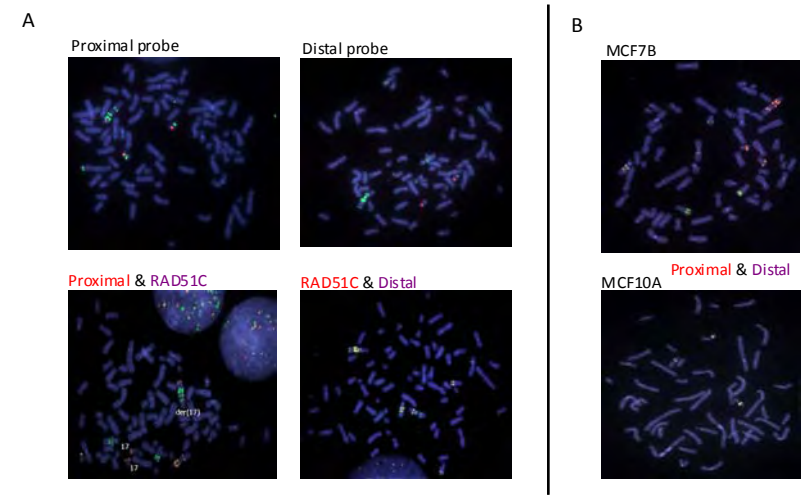

Figure 9: Detection RAD51C/ATXN7 fusion by FISH. A) Proximal and distal probes were tested on metaphase spreads of MCF-7 cells. MCF-7 shows clear amplification of RAD51C. B) Break-away FISH on metaphases of MCF-7 and MCF-10A cells.

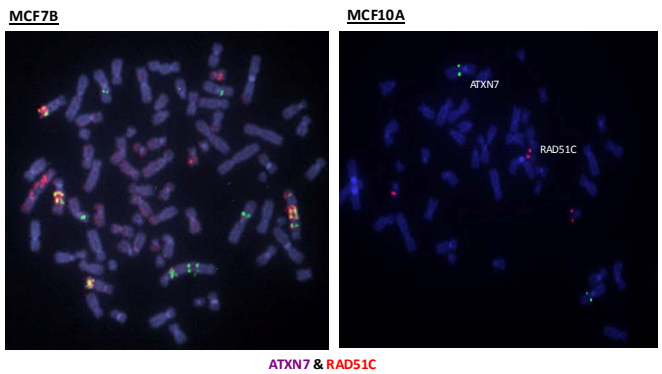

Figure 10: Detection RAD51C/ATXN7 fusion by FISH. Metaphases of MCF-7 and MCF10A cells were hybridized with a RAD51C probe and ATXN7 probe. Yellow signal in MCF-7 shows co-localization of DA51C and ATXN7, indicating a fusion between the genes.

b. Detect recurrent breaks with break-away FISH and associate with histo-pathological and clinical characteristics.

We were unable to validate the presence of the fusion on genomic level by FISH in T47D, and MDA MB 361. Because of our inability to validate the genomic fusion by FISH, and the fact that this technique is very low-throughput, we decided to focus our attention and resources on developing a high-throughput screening technique. We also used these data for the re-discovery of break-points and for the discovery of previously unidentified breakpoints.

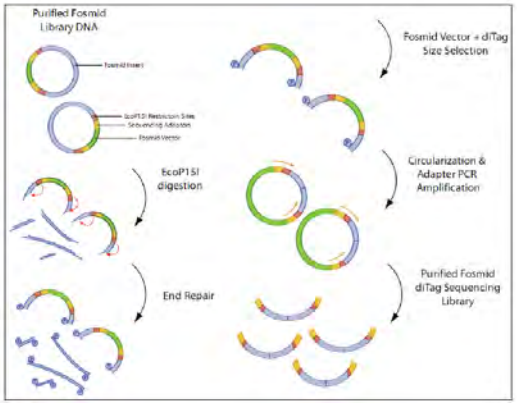

Figure 11: Fosmid diTag workflow schema illustrating the formation and amplification of the concatenated 26bp end-tags from the fosmid insert termini.

Here we were faced with 2 key challenges: (1) sensitive detection of chromosomal aberrations, and (2) high-throughput validation of putative chromosomal aberrations. To address both challenges, we have developed the fosmid diTag method, the first long-range mate-pair physical mapping method based on massively parallel sequencing. We apply the fosmid diTag and 5 Kbp Illumina mate pair methods to MCF7 and HCC1954 breast cancer cell lines (Figure 11). The rearrangements detected by both methods show a 3-fold enrichment for cancer-specific somatic mutations compared to those detected by the 5Kbp method alone. Fosmid diTag method also reveals much higher proportion of gene fusions and truncations. We first mapped aberrations in MCF7 and HCC1954 using the medium-range Illumina mate pair method. To detect rearrangements, we searched for regions where at least two independent pairs of ends showed discrepancy with their predicted size and/or orientation. We found 23,5555 rearrangements in MCF7 and 3,824 in HCC1954. A total of 18,444 indels were identified in MCF7 and 1,258 in HCC1954 (Figure 12). Other rearrangement classes are based on incorrect orientation, indicative of potential inversions, or whose ends map to different chromosomes, predicting an interchromosomal translocation. In MCF7, we detected 1,575 inversions and 3,536 translocations. We found 485 inversions and 2,081 translocations in HCC1954 (Figure 12).Again we were intested in the validation rate by PCR, and primers spanning the breakpoints were tested on genomic DNA from breast cancer cell lines and normal controls. The PCR assay produced a single strong amplification product in 28% and 37% of the reactions for MCF7 and HCC1954, respectively. We detect and validate a total of 91 somatic rearrangements in MCF7 and 25 in HCC1954, including genomic alterations corresponding to 50% of the transcript fusions previously discovered by transcript mapping.

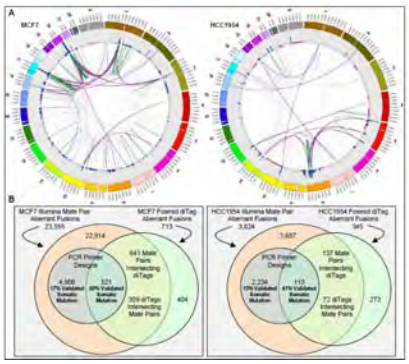

Figure 12: (A) Circular visualizations of the MCF7 and HCC1954 genomes obtained using Circos (Krzywinski, Schein et al. 2009) software. (B) Venn diagrams comparing the number of intersecting aberrant fusions, PCR primer designs, and somatic mutation validation between fosmid diTag and Illumina mate pair data sets in the MCF7 and HCC1954 genomes.

Task 4: *Study the biological significance of the breakpoints using in vitro models.*

a. Determine the downstream consequence based on the position of the aberrant joint.

Based on protein sequence analysis and protein translation programs, I was able to predict the fusion protein, and speculate on the consequence of the ARFGEF2/SULF2, and RAD51C/ATXN7. By the creation of the ARFGEF2/SULF2 fusion, the Sulfatase 2 (SULF2) protein loses its targeting peptide for targeting for secretion, while the added sequence of ARFGEF2 does not add any functional domain. This might mean that the fusion creates a non-functional Sulfatase 2.
By using protein translation programs, it became clear that with the creation of the fusion between RAD51C and ATXN7 a frameshift in the codon is induced. This translates into the introduction of a stop-codon early in the ATXN7 sequence. This most likely results in the loss of a critical C-terminal domain of RAD51C, without the addition of any significant sequence of Ataxin 7. Preliminary data confirming this truncation is shown in Task 1.

b. Recreate join with cloning techniques.

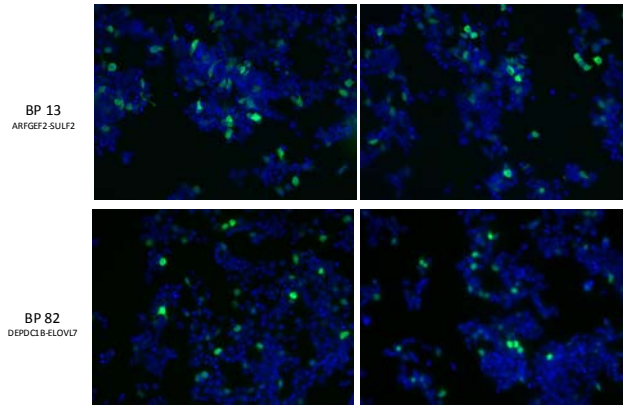

Figure 13: Test expression of fusion constructs. 293 cells were transfected with either control plasmid (not shown), ARFGEF2/SULF2, or DEPDC1B/ELOVL7 constructs. Cells were fixed, stained with an anti-V5 antibody, and imaged by fluorescence microscopy.

I have cloned ARFGEF2/SULF2, DEPDC1B/ELOVL7 and RAD51C/ATXN7 fusions into mammalian expression vectors by performing RT-PCR on MCF-7 cells. The expression of the ARFGEF2/SULF2, and the DEPDC1B/ELOVL7 vectors have been tested by transfecting 293 cells. The cells were then stained with an anti-V5 antibody, and analyzed by fluorescence (Figure 13).

c. Perform targeted experiments to determine functional consequence of the aberrant join.

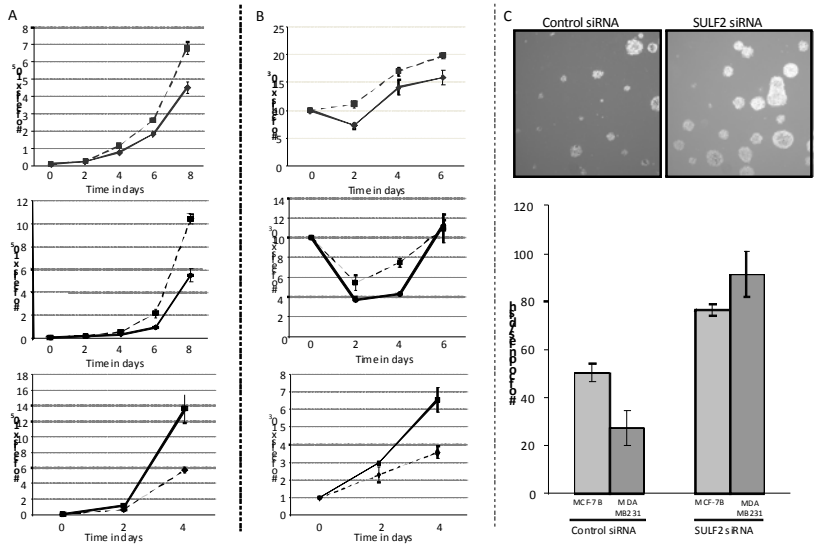

Figure 14: A) Cells treated with *SULF2* siRNA have an enhanced proliferation then cells treated with control siRNA. B) Cells treated with *SULF2* siRNA have an enhanced survival compared to cells treated with control siRNA. C) Treatment of MCF-7B and MDA MB231 cells with siRNA for *SULF2* increases the anchorage-independent growth capabilities.

To give insight into the function of the ARFGEF2/SULF2 fusion, SULF2 mRNA was knocked down using siRNA specifically targeting SULF2 in MCF-7, MDA MB231 and MCF10A cells. All three cell lines treated with SULF2 siRNA used in a proliferation assay, exhibited an advantage over the cells treated with control siRNA. Also, cells treated with SULF2 siRNA showed an enhanced survival. Cells with reduced SULF2 die less, and recover faster in

serum free conditions than control cells. Knock-down of SULF2 mRNA also gave a clear advantage in anchorage-independent growth capability. This shows that knocking-down SULF2 enhances the tumorigenic properties in multiple breast cell lines, and that SULF2 might act as a tumor-suppressor in breast cancer development. The presence of this ARFGEF2/SULF2 fusion might mean a loss of function of the wildtype tumor suppressor Sulfatase 2 and enhance the tumorigenicity of MCF-7 cells.

Key research Accomplishments

- Discovered 157 joins in MCF-7 cell line, of which only few have been previously described.
- 10 gene fusion were discovered, of which 4 express a chimeric mRNA (3 new ARFGEF2/SULF2, 1 previously described)
- 31 of the 157 are present in all 7 MCF-7 sublines tested. This allows us to narrow down on 'true' breakpoints present in the ancestor of the MCF-7 cell lines.
- Confirmed RAD51C/ATXN7 fusion by FISH in MCF-7 cell line.
- Cloned 3 fusion transcripts (ARFGEF2/SULF2, DEPDC1B/ELOVL7, RAD51C/ATXN7) into mammalian expression vectors by amplification of the fusion transcript by RT-PCR.
- Discovery of RAD51C/ATXN7 fusion transcript in two other breast cancer cell lines (T47D, and MDA MB361)
- Discovery of short form of Rad51C protein in MCF-7 and MDA MB361
- Sulfatase 2 acts as a tumor suppressor in breast cancer cell lines, and might be dysfunctional after generation of the ARFGEF2/SULF2 fusion.

Training accomplishments

- Presented twice at the Research and Development workshop of the Breast Center.
- Attended and presented orally data at the Breast Center/Cancer Center retreat (November 2008)
- Attended and presented a poster at the LINK meeting (February 2009)
- Attended and presented a poster at the Breast Center/Cancer Center retreat (September 2009)
- Attended weekly the Research and Development workshop of the Breast Center
- Attended bi-monthly the Journal Club of the Breast Center
- Contributed to the generation of data, writing and editing of the manuscript published in Genome Research
- Attended the course 'Translational Breast Cancer'
- Supervised several graduate and summer students.

Reportable outcomes

- Hampton OA, **den Hollander P**, Miller CA, Delgado DA, Li J, Coarfa C, Harris RA, Richards S, Scherer SE, Muzny DM, Gibbs RA, Lee AV, Milosavljevic A: A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Research*. Feb;19(2):167-77 2009.
- Abstract submission for the Breast Center Retreat November 2008 entitled: Discovery of functional genomic breakpoints in breast cancer.

- Abstract submission for the Breast Center Retreat September 2009 entitled: Evolution of genomic diversity in the breast cancer cell line MCF-7.
- Abstract submission the San Antonio Breast Cancer Symposium December 2009 entitled: Evolution of genomic diversity in the breast cancer cell line MCF-7
- Oliver A. Hampton, Christopher A. Miller, Maxim Koriabine, Jian Li, **Petra Den Hollander**, Lucia Carbone, Mikhail Nefedov, Boudewijn F.H. Ten Hallers, Adrian V. Lee, Pieter J. De Jong, Aleksandar Milosavljevic: Long-range massively parallel mate pair sequencing detects distinct mutations and similar patterns of structural mutability in two breast cancer cell lines. *Cancer Genetics*. 204: 447-457 2011
- Evolution of genomic diversity in the breast cancer cell line MCF-7. *In preparation*.

Personnel receiving pay from the research effort

- Petra den Hollander, PhD

Conclusion

In contrast to leukemias and lymphomas, carcinomas contain more complex chromosomal rearrangements, only partially detectable using classic cytogenetic methods. Thus, our knowledge of chromosomal rearrangements in solid tumors is very limited, and "gene fusions" defining a specific type of solid tumor have not yet been characterized. This lack of knowledge has supported the paradigm that chromosomal rearrangements leading to gene fusions are almost exclusively seen in hematologic malignancies and are extremely rare (maybe <1%) in solid tumors.

Here we set out to discover the chromosomal rearrangements that are important in breast cancer. The data presented here shows that there are indeed breakpoints that have a functional significance in breast cancer cell lines. I even discovered a fusion that is present in two other breast cancer cell lines besides MCF-7. The data shown on the ARFGEF2/SULF2 and RAD51C/ATXN7 fusions indicate that we discovered novel strategy of the tumor cells to silence important tumor suppressors.

To gain insight into the heterogeneity of genomic breakpoints, in seven MCF-7 sub-lines. There is an enrichment for breakpoints containing genes (50.3% vs 77.4%), and for fusion-containing breakpoints (6.4% vs 16.1%). When studying cell lines originating from a single cell, we discovered that there is very little genetic variability between them. A large effort has gone into the development of next-generation sequencing techniques for the discovery of genomic breaks and fusion in the breast cancer genome. We developed new techniques and validated them with standard PCR. The validation rate is very promising, and these new techniques will aid us in the discovery of breast cancer.

A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome

Oliver A. Hampton, Petra Den Hollander, Christopher A. Miller, et al.

Genome Res. 2009 19: 167-177 originally published online December 3, 2008
Access the most recent version at doi:10.1101/gr.080259.108

Letter

A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome

Oliver A. Hampton,[1,3,5] Petra Den Hollander,[4,5] Christopher A. Miller,[1,3] David A. Delgado,[4,5] Jian Li,[1,3] Cristian Coarfa,[1,2] Ronald A. Harris,[1,2] Stephen Richards,[2] Steven E. Scherer,[2] Donna M. Muzny,[2] Richard A. Gibbs,[2,3] Adrian V. Lee,[4,5,6] and Aleksandar Milosavljevic[1,2,3,5,7]

[1]*Bioinformatics Research Laboratory, Baylor College of Medicine, Houston, Texas 77030, USA;* [2]*Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA;* [3]*Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas 77030, USA;* [4]*Breast Center, Baylor College of Medicine, Houston, Texas 77030, USA;* [5]*Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas 77030, USA;* [6]*Department of Medicine, Baylor College of Medicine, Houston, Texas 77030, USA*

By applying a method that combines end-sequence profiling and massively parallel sequencing, we obtained a sequence-level map of chromosomal aberrations in the genome of the MCF-7 breast cancer cell line. A total of 157 distinct somatic breakpoints of two distinct types, dispersed and clustered, were identified. A total of 89 breakpoints are evenly dispersed across the genome. A majority of dispersed breakpoints are in regions of low copy repeats (LCRs), indicating a possible role for LCRs in chromosome breakage. The remaining 68 breakpoints form four distinct clusters of closely spaced breakpoints that coincide with the four highly amplified regions in MCF-7 detected by array CGH located in the 1p13.1-p21.1, 3p14.1-p14.2, 17q22-q24.3, and 20q12-q13.33 chromosomal cytobands. The clustered breakpoints are not significantly associated with LCRs. Sequences flanking most (95%) breakpoint junctions are consistent with double-stranded DNA break repair by nonhomologous end-joining or template switching. A total of 79 known or predicted genes are involved in rearrangement events, including 10 fusions of coding exons from different genes and 77 other rearrangements. Four fusions result in novel expressed chimeric mRNA transcripts. One of the four expressed fusion products (*RAD51C–ATXN7*) and one gene truncation (*BRIP1* or *BACH1*) involve genes coding for members of protein complexes responsible for homology-driven repair of double-stranded DNA breaks. Another one of the four expressed fusion products (*ARFGEF2–SULF2*) involves SULF2, a regulator of cell growth and angiogenesis. We show that knock-down of *SULF2* in cell lines causes tumorigenic phenotypes, including increased proliferation, enhanced survival, and increased anchorage-independent growth.

[Supplemental material is available online at www.genome.org and through the Breast Cancer project page at www.genboree.org. All MCF-7 BAC clones are available from Amplicon Express under name HTA and plate/row/column names as indicated. The sequence data from this study have been submitted to the NCBI Trace and Short Read Archives (http://www.ncbi.nlm.nih.gov) under accession nos. 2172834909–2172901416 and 2172904832–2172911164, and SRR006762–SRR006767, respectively].

Many cancer genomes are characterized by mutability, including microsatellite instability (MIN) and chromosomal instability (CIN) (Lengauer et al. 1998). It is now generally anticipated that sequencing of cancer genomes using massively parallel sequencing technologies (Korbel et al. 2007; Campbell et al. 2008) will provide insights into structural mutability. Recent sequencing of four cancer amplicons (Bignell et al. 2007) derived from the HCC1954 breast cancer cell line and two lung cancer cell lines provided evidence for homologous and nonhomologous repair of double-strand DNA breaks induced by the breakage-fusion-bridge (BFB) mechanism.

[7]**Corresponding author.**
E-mail amilosav@bcm.edu; fax (713) 798-4373.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.080259.108. Freely available online through the *Genome Research* Open Access option.

Gene fusions and truncations that result from chromosomal rearrangements provide insight into the molecular mechanisms of cancer progression. Recurrent rearrangements of specific genes indicate increased mutability or positive selection (or a combination of both) in the evolution of tumor genomes. Recurrent fusions, translocations, and other aberrant joins are used as highly informative diagnostic and prognostic markers and drug targets in leukemias, lymphomas, and sarcomas. A total of 337 genes involved in fusions in cancer genomes have been recently surveyed (Mitelman et al. 2007). Four gene fusions have previously been reported in breast carcinomas (*ETV6–NTRK3*, *ODZ4–NRG1*, *TBL1XR1–RGS17*, *BCAS3-BCAS4*) (Mitelman et al. 2007, Ruan et al. 2007).

Breast cancer and carcinomas in general have proven less tractable to fusion discovery due to the typically higher degree of rearrangement. However, a prognostically significant rearrangement was recently discovered in the majority of prostate cancers (Tomlins et al. 2005). Of note, the initial discovery was not iden-

tified by analyzing DNA sequence or structure, but via the analysis of outlier gene expression, followed by a targeted locus-specific search for a fusion in genomic DNA. Here we demonstrate a method to detect gene fusions directly by the analysis of genomic DNA, even in highly rearranged breast cancer.

MCF-7 is the most widely used cell line model for estrogen-positive breast cancer. The cell line has been derived from a pleural effusion taken from a patient with metastatic breast carcinoma (Soule et al. 1973). Evidence of CIN in MCF-7 comes from apparent aneuploidy and significant genomic divergence in several sublines (Jones et al. 2000; Nugoli et al. 2003). Chromosomal aberrations in MCF-7 have previously been studied by spectral karyotyping (Kytola et al. 2000; Rummukainen et al. 2001), comparative genomic hybridization (CGH) (Kytola et al. 2000; Rummukainen et al. 2001), array CGH (Neve et al. 2006; Shadeo and Lam 2006; Jonsson et al. 2007), single nucleotide polymorphism arrays (Huang et al. 2004), and gene expression arrays (Neve et al. 2006).

More recently, bacterial artificial chromosome (BAC)-based end sequence profiling (ESP) (Volik et al. 2003, 2006; Raphael et al. 2008) has been applied to study genomic rearrangements in cancer genomes. Volik and colleagues sequenced a total of 19,831 BAC ends from the Amplicon Express MCF-7 BAC library, ~1× clone coverage of the human genome, to identify 582 BACs containing rearrangements.

As a starting point for our analysis, we constructed BAC pools from a nonredundant subset (n = 552) of rearranged BACs identified by Volik et al. (2003, 2006). To map chromosomal aberrations in the genome of the MCF-7 breast cancer cell line at sequence level resolution, we developed a method that combines end-sequence profiling and massively parallel sequencing. By analyzing sequences of the chromosomal breakpoints in the BAC pools, we gained insights into the mechanisms of chromosomal instability and repair. Specific gene fusions and truncations that have emerged during the pathological evolution of this cancer genome point to the molecular mechanisms of the disease. Additional products of our research are benchmarking reagents for the development of a new generation of methods for detecting structural genome variation, including well-characterized BAC pools and validated breakpoints in the MCF-7 genome.

Results

At least 157 breakpoints were induced by somatic rearrangements in MCF-7

Aberrant breakpoint-induced joins were identified by combining "bridging" and "outlining" steps, as illustrated in Figure 1A. The bridging step utilizes end-sequence information from fosmid-sized clone inserts to connect chromosomal loci brought together at aberrant rearrangement-induced joins in the cancer genome. End-sequences of breakpoint-spanning fosmids were recognized as those that do not map onto the reference genome in a manner consistent with the clone insert size or end-sequence orientation. The outlining step involves a precise localization of breakpoint sites by mapping short tags generated by the 454 Life Sciences (Roche) pyrosequencing machine onto the reference genome.

As illustrated in Figure 1B, three pools, each containing 192 BACs containing putative rearrangements, were constructed for the purpose of massively parallel sequencing using the 454 GS sequencing machine. Approximately 300,000 short (~100-bp) reads were sequenced from each pool, providing ~1× sequence coverage for the purpose of outlining. Six 96-BAC pools were formed from the same set of BACs for the purpose of fosmid library preparation, end-sequencing and bridging. Approximately 8000 to 10,000 fosmid inserts from each of the six pools were end-sequenced, providing 24× clone coverage and ~1× sequence coverage for the purpose of bridging.

Upon sequencing, the fosmid end-reads and the 454 reads together with the BAC end-sequences produced by Volik et al. (2003, 2006) were mapped onto the reference human genome. Independent aberrant mapping of two fosmids across a specific putative breakpoint was considered to constitute sufficient evidence to declare the breakpoint. BAC or fosmid ends that map onto different chromosomes are interpreted as interchromosomal breakpoints. The outlined regions were bridged using end-sequences from BACs and fosmids. The combination of outlining and bridging enabled identification of breakpoint locations down to a PCR-able distance. As indicated in Figure 1C, out of the total of 410 detected breakpoints, 157 could be confirmed by PCR across breakpoint joins as likely distinct somatic mutations. As indicated by the bars in the middle of Figure 1C, the remaining breakpoints failed the confirmation process for a number of different reasons, as we explain next.

A total of 47 breakpoints could not be unambiguously resolved down to a PCR-able distance using the outlining method. PCR primers were designed for the remaining breakpoints using a semi-automated primer design pipeline. When applied to pooled BACs, PCR primers failed to generate amplicons in expected size range for 23 predicted breakpoint joins. Further confirmation included amplification of a pool of genomic DNA from six MCF-7 cell lines (B, BK, C, D, L, and Neo). DNA isolated from MCF-10A and normal human female DNA (Novagen) were used as negative controls. A total of 123 PCR primer pairs that produced amplicons from the BAC pool did not produce amplicons from the genomic DNA derived from cell pools. A majority of these breakpoint sites contained HindIII restriction sites. Since the BAC library was prepared using HindIII partial-digestion of genomic DNA, those breakpoints were most likely created by fusion of digestion products in the course of BAC library preparation. Other sources of this discrepancy may include a number of cell line–specific aberrations generated over a number of passages that preceded preparation of the BAC library.

To identify structural polymorphic variants present in the germline of the MCF-7 donor, PCR amplification of breakpoint joins was performed on a pool of 90 Caucasian HapMap genomes (International HapMap Consortium 2005). Additionally, search for occurrences of the apparently somatic joins was performed in publicly available genomic sequences using the Pash program (Kalafus et al. 2004). A total of 40 apparently aberrant joins were present in the HapMap samples, as indicated by the presence of a PCR product, and thus correspond to structural alleles different from the structural alleles represented in the reference genome assembly. Finally, some breakpoints were identified to occur in more than one BAC, and the count was reduced by 20 to eliminate multiple counting, resulting in a total of 157 unique confirmed somatic breakpoint joins in the MCF-7 genome. Of the 157 MCF-7 somatic breast cancer breakpoints, 74 (47%) formed interchromosomal and 83 (53%) intrachromosomal joins, as illustrated in Figure 2, A and B.

A majority of the somatic breakpoints could be assigned to specific BACs

If a chromosomal segment outlined by 454 reads connected a BAC end-sequence and a breakpoint-spanning fosmid end-sequence,

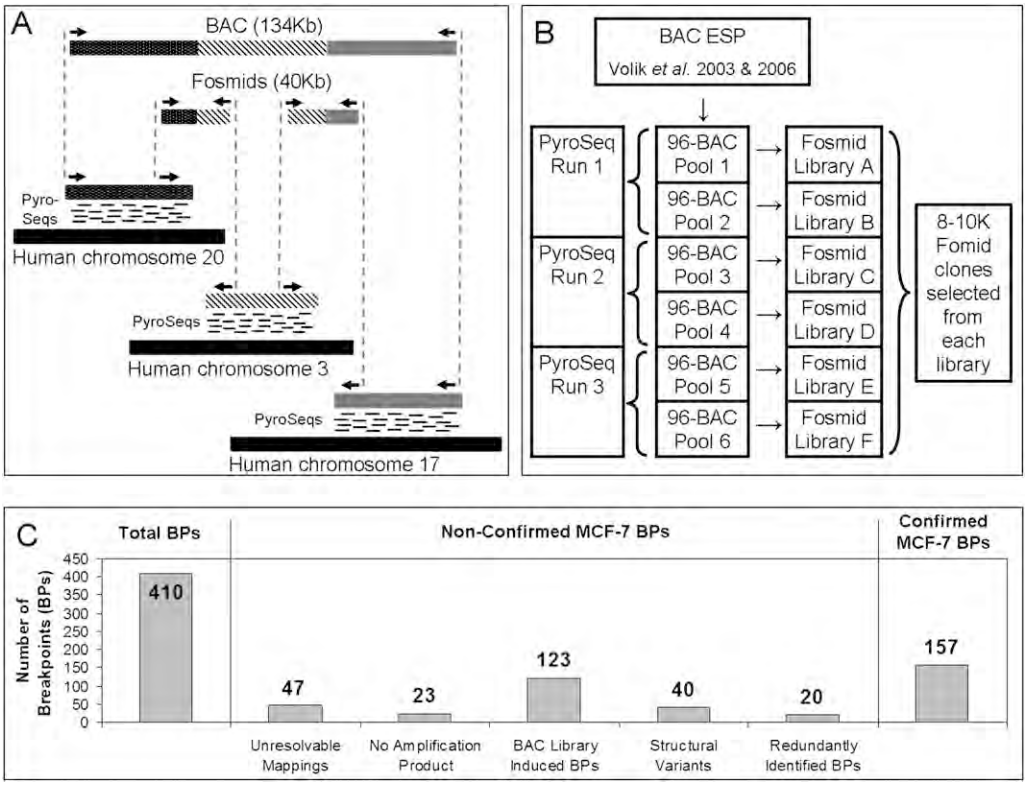

Figure 1. (*A*) An illustration of the principle of the method. Breakpoints within a BAC containing segments from chromosomes 20, 3, and 17 are detected using a combination of "bridging" and "outlining" steps. The bridging step maps fosmid end-sequences onto the reference genome. The outlining step maps short tags (labeled "PyroSeqs") using 454 technology from the BAC (in practice a pool of BACs) onto the reference genome. The results of bridging and outlining jointly allow precise mapping of breakpoints and reconstruction of rearranged BACs. (*B*) Organization of the mapping experiment. The nonredundant collection of 552 rearrangement containing BACs, 17 normal BAC negative controls, and seven positive controls was arrayed in six 96-well plates and pooled as indicated. Three 454 sequencing reactions (involving BACs pooled from plate pairs) produced tags for the purpose of outlining. Six fosmid libraries (one from each 96-well plate pool of BACs) were constructed for Sanger-based sequencing of fosmid ends and bridging. (*C*) Bar charts detailing the classification of detected MCF-7 breakpoints.

the breakpoint could be associated with the BAC. Out of 552 pooled BACs, at least one breakpoint could be assigned to 316 (57%) of them. The remaining BACs fall into the following two groups: First, in 129 (23%) cases, breakpoint assignment was inconclusive due to ambiguous mapping of reads onto the reference genome, mostly due to repetitive DNA regions, apparent overlaps between BACs, and other causes; second, in 107 (20%) cases, a single outlining block connected BAC ends, thus indicating lack of any rearrangement, contrary to previous reports (Volik et al. 2003, 2006).

To examine the source of the disagreement with the previous reports, the 107 disagreements were examined in detail. Most of the disagreements could be explained either by the differences between reference genome assemblies used in the previous and current studies or by mismapping of BAC-end sequence reads or by a combination of the two factors. Assemblies used in the previous studies were NCBI Build 30 of June 2002 (Volik et al. 2003) and NCBI Build 34 of July 2003 (Volik et al. 2006), while our study employed NCBI Build 36 of March 2006. The newer assembly is more likely to be more correct and complete, but some of the disagreements may also be explained by the presence of different structural alleles at sites of structural polymorphisms. The disagreements tended to occur in regions containing low copy repeats (LCRs). For example, Volik et al. (2003) identified MCF-7 BAC 9I10 as bridging apparent translocation t(11;11)(p11.12;q14.3) and apparently confirmed the rearrangement by fluorescent in situ hybridization (FISH). Examination of Build 36 reveals copies of an LCR at both 11p11.12 and 11q14.3. The LCR was absent from Builds 30 and 34, thus explaining the aberrant BAC-end sequence mapping and even the erroneous "confirmation" by FISH.

Examination of breakpoint sequences reveals signatures of DSB repair

To examine breakpoints at the sequence level, all the 157 breakpoint-spanning amplicons were used as substrates for sequencing from both ends. Most amplicons were of small enough size (less than 1 kb on average), allowing the Sanger read from at least one of the ends to reach the breakpoint. Difficultly of sequencing across breakpoints has been documented (Lee et al. 2007; Liu and Carson 2007), especially in repeat-rich regions. To ameliorate the problem, we sequenced DNA from specific BAC pools and employed nested sequencing primers in cases of first-pass sequencing failures. Breakpoint-straddling sequence could be obtained from 86 (55%) amplicons and could not be obtained for the remaining 71 (45%). Many of the failures were due to inability to design unique primers for sequencing across breakpoints that fall within repeat-rich regions.

Examination of 86 breakpoints that could be resolved to the base pair level (summarized in the chart in the middle of Fig. 2B)

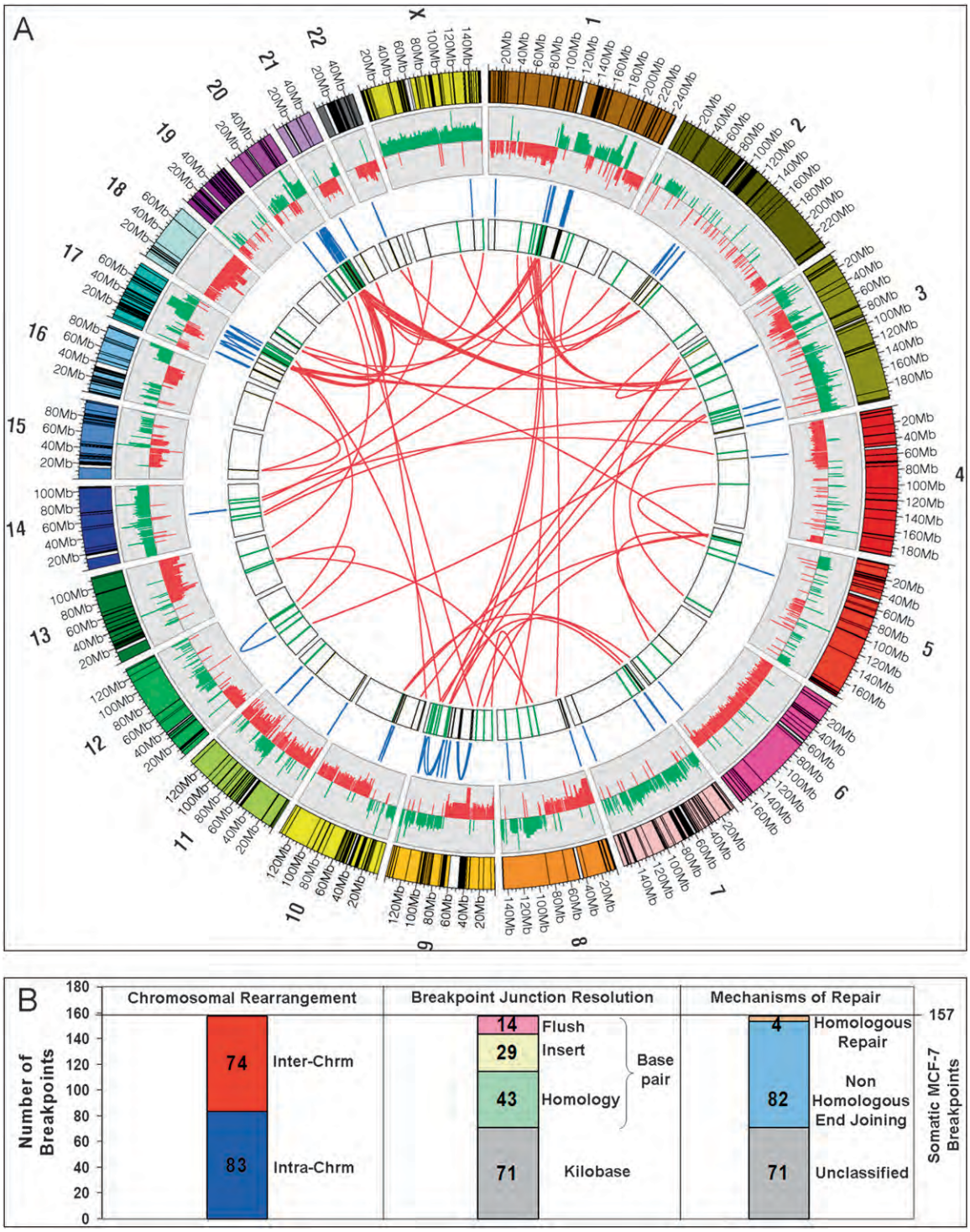

Figure 2. (*A*) Circular visualization of the MCF-7 genome obtained using Circos software. Chromosomes are individually colored with centromeres in white and LCR regions in black. MCF-7 BAC array comparative genome hybridization data (Jonsson et al. 2007) are plotted with gains in green and losses in red using $\log_2$ratio. The inner chromosome annotations depict 157 somatic MCF-7 breast tumor chromosomal rearrangements associated with LCRs (black) and breakpoints not associated with LCRs (green). Chromosomal rearrangements are depicted on each side of the MCF-7 breakpoints; intra-chromosomal rearrangements (blue) are located outside and interchromosomal rearrangements (red) are located in the center of the circle. (*B*) Bar charts indicating classification of somatic breakpoints in MCF-7.

revealed 14 flush joins without evidence of microhomology or intervening sequence, 29 joins with intervening inserts of unknown genomic origin averaging over 100 bp in length, and 43 joins where the joined segments exhibit homology. The extent of homology was in most (88%) cases restricted to ≤7 bp, consistent with microhomology observed in double-stranded breaks repaired by nonhomologous end-joining (NHEJ) or template switching (Sonoda et al. 2006). Due to the absence of straddling sequence, the remaining 71 breakpoints could only be analyzed at the ~1-kbp level of resolution.

Out of the 86 somatic breakpoints isolated to base pair resolution, only four (5%) exhibited sequence patterns—sequence identity and equal crossover between two homologous loci—consistent with nonallelic homologous recombination (NAHR) (chart on the right of Fig. 2B). The dominant mechanism responsible for the repair of double-strand breaks in MCF-7 therefore appears to be NHEJ or template switching.

Two distinct types of breakpoints exist in MCF-7–clustered and LCR-associated

As evident from Figure 2, the breakpoints in MCF-7 are not evenly distributed across the genome. A number of clusters of closely spaced breakpoints are evident. To formally delineate the clustered breakpoints from the remainder, clusters of eight or more breakpoints that are less than 1.1 Mbp apart were identified. Four such clusters emerged in the following locations: 1p13.1-21.1, 3p14.1-p14.2, 17q22-q24.3, and 20q12-q13.33. These four rearrangement clusters, illustrated in Figure 3A, contain 43% of all MCF-7 somatic breakpoints, while representing only 1.5% of the normal reference genome.

The remaining nonclustered or dispersed breakpoints are highly associated with LCRs, showing a 5.2-fold enrichment for the presence of LCRs at the breakpoint site (P-value = 2.9×10^{-22}; see Fig. 3B). This is in contrast to the clustered breakpoints that do not exhibit enrichment for LCRs, with only five out of 68 clustered breakpoints being LCR-associated, well within the number expected by chance. Moreover, as illustrated in Figure 3C, the four clustered breakpoint locations exactly coincide with high copy number gain regions ("firestorms," the term proposed by Hicks et al. [2006]) in the MCF-7 genome described by Jonsson et al. (2007) and contain prognostic gene markers for breast cancer.

To further examine possible differences between the clustered breakpoints and the dispersed ones, we identified regions that show recurrent copy number amplification in cancer in previous studies involving 145 breast tumors and 56 breast cancer cell lines (Chin et al. 2006; Neve et al. 2006; Shadeo and Lam 2006; Jonsson et al. 2007). As illustrated in Supplemental Figure 5, almost three-fourths of breakpoints occurring in the four clusters are highly recurrently amplified (high recurrence is declared if at least 20% of the surveyed samples show amplification), a greater than twofold enrichment over other (dispersed) breakpoints. Additionally, the mean number of amplifications at each breakpoint location is significantly higher among clustered vs. dispersed breakpoints. These data suggest that genomic instability in these cluster regions is not specific to MCF-7.

Novel chimeiric transcripts could be predicted based on fusions of genomic DNA

Among the breakpoint fusions that involved genes, we first focused on those that occurred within introns and are predicted to lead to chimeric transcripts. We discovered 10 gene fusions (Table 1) where fusion breakpoints reside in intronic regions of the genes involved, implying in-frame translation of the original amino acid sequences.

To determine if the predicted chimeric mRNA transcript was created by these genomic fusions, we performed gene-specific reverse transcriptase reactions and a fusion-specific PCR on RNA extracted from MCF-7, MCF-10A, and normal breast tissue (the latter two serving as negative controls). Since the primers were designed to amplify the fusion product specifically, a band was only generated if a fusion product was present (for primers sequence see Supplemental Table 4). Out of 10 fusions, four showed a fusion mRNA transcript by RT-PCR, see Figure 4.

To identify if other sources reported the same fusion transcripts in MCF-7, other cell lines or primary tumors, we queried 70 MCF-7 and HCT116 (colon cancer) paired-end ditag fusion transcript sets reported by Ruan et al. (2007) and 237 fusion transcripts from the Cancer Genome Anatomy Project Recurrent Chromosome Aberrations in Cancer database reported by Hahn et al. (2004). Of the 10 MCF-7 gene fusions identified by our bridging and outlining method, the *BCAS3-BCAS4* fusion was found to be previously characterized Ruan et al. (2007) Interestingly, the *BCAS3-BCAS4* fusion is recurrently present in both the MCF-7 breast cancer and HCT116 colon cancer cell lines.

Some of the fusions and truncations may suppress function of normal gene product

Most fusions involve highly amplified clustered breakpoints, indicating possible positive selection and therefore functional significance. This is consistent with the fact that firestorm patterns indicate poor prognosis (Hicks et al. 2006) and that these highly amplified regions contain specific prognostic markers (Jonsson et al. 2007). However, not all the amplified loci contain oncogenes. Analysis and results below indicate that the oncogenic effects of some of the fusions may in fact be due to a suppression of normal function of a tumor suppressor gene. Observed amplification of gene fusions involving tumor suppressors is consistent with a dominant-negative effect of such gene fusions.

For example, the first two exons of *PTPRG*, comprising the carbonic anhydrase-like domain, are replaced by the first 10 exons of the unannotated inter-species *ASTN2* gene. Promoter hypermethylation in *PTPRG* in T-cell lymphoma leads to loss of gene expression and correlates with poor prognosis (van Doorn et al. 2005). Interestingly, Murine L cells producing *PTPRG* transcripts with a homozygous deletion of the carbonic anhydrase-like domain causes sarcomas in syngeneic mice (Wary et al. 1993).

To examine the effects of a possible suppression of *SULF2* function by the *ARFGEF2-SULF2* fusion, *SULF2* mRNA was knocked down using siRNA specifically targeting *SULF2* in MCF-7B, MDA MB231, and MCF-10A cells (Supplemental Fig. 6). Proliferation assays were performed on the three cell lines treated with knocked down *SULF2*, and all exhibited an advantage over the cells treated with control siRNA (Fig. 5A–C). To determine the effect on survival capabilities under stress conditions, *SULF2* siRNA and control siRNA treated cells were plated in serum-free conditions. Results indicate (Fig. 5D–F) that cells with knocked down *SULF2* survive better, and recover faster (seen by the steeper slope) in serum-free conditions then the control cells. This implies that knock-down of *SULF2* enhances survival compared to the control cells. Finally, knock-down of *SULF2* mRNA caused a twofold increase in anchorage-independent growth in MCF-7B and a threefold increase

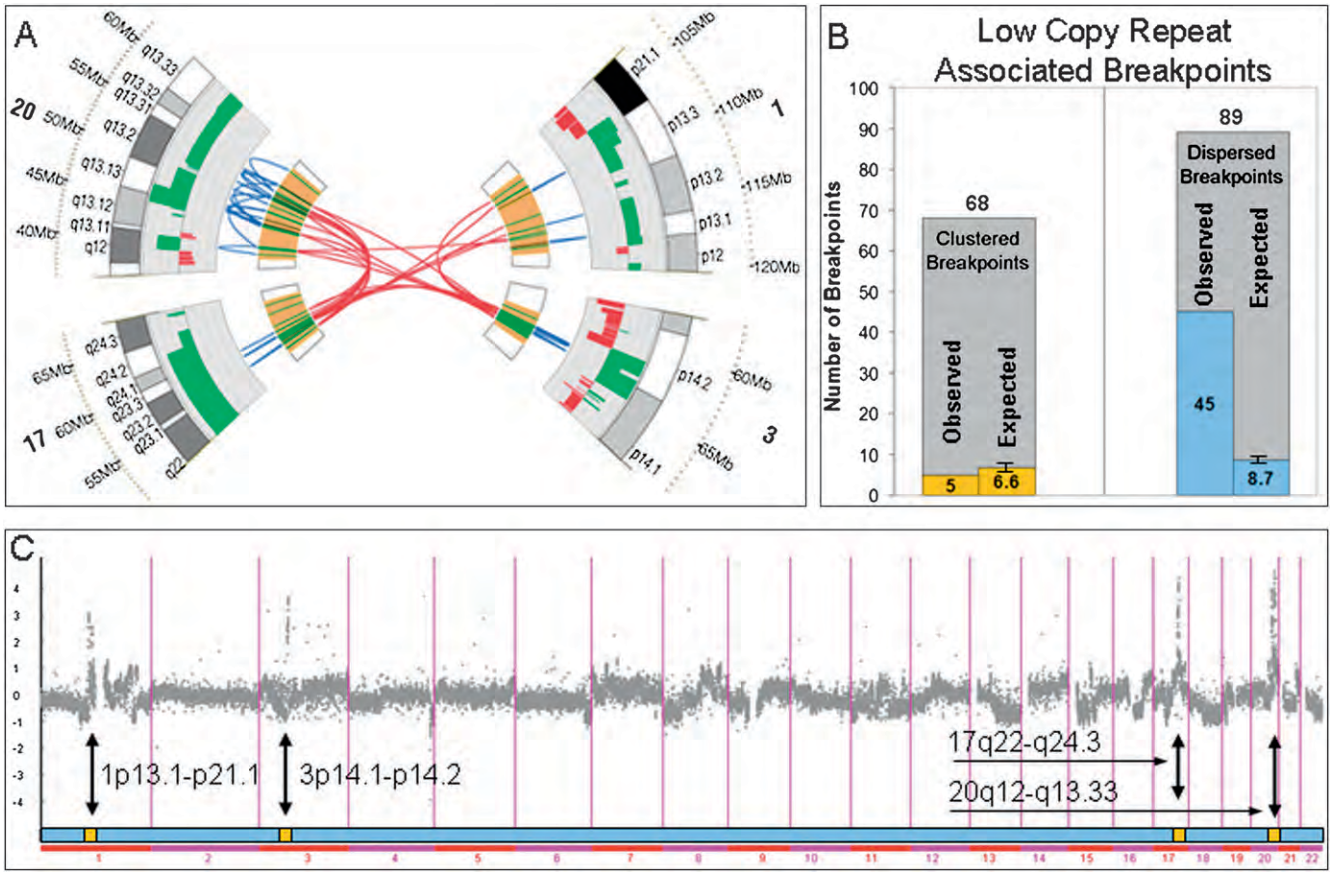

Figure 3. (*A*) Four clusters of breakpoints at 1p13.1-21.1, 3p14.1-p14.2, 17q22-q24.3, and 20q12-q13.33. (*B*) Low copy repeat (LCR) association with clustered and dispersed breakpoints. (*C*) The four clusters of breakpoints correspond exactly to the four highly amplified regions in MCF-7, as determined by array CGH.

in MDA MB231, as measured by the amount of colonies compared with controls (Fig. 5H). In summary, the data indicate that knockdown of *SULF2* causes tumorigenic phenotypes, including increased proliferation, enhanced survival, and increased anchorage-independent growth. *SULF2* may therefore act as a breast cancer suppressor.

Some genes are involved in numerous rearrangements

In addition to the 10 gene–gene fusions, a total of 77 genes were otherwise affected by the 157 breakpoints. We jointly refer to those events as "truncations" even though some, in fact, involve fusion of an upstream promoter with a protein coding gene. *PTPRG* and other genes were affected by multiple breakpoints, including both

Table 1. Gene fusions in MCF-7 that involve splicing of intact coding exons

Associated genes	Rearrangement type	Cytoband translocation	Comment
ARFGEF2-SULF2	Intrachromosomal inversion	20q13.13-20q13.13	Fusion of *ARFGEF2* exon 1 to *SULF2* exons 3–21; 1.2-Mb inversion
DEPDC1B-ELOVL7	Intrachromosomal translocation	5q12.1-5q12.1	Fusion of *DEPDC1B* N terminus exons 1–7 (out of 11) with *ELOVL7* exons 8–9
RAD51C-ATXN7	Interchromosomal rearrangement	3p14.1-17q22	Fusion of *RAD51C* exons 1–7 (out of nine) with *ATXN7* exons 6–13
SULF2-PRICKLE2	Interchromosomal rearrangement	3p14.1-20q13.13	Fusion of *SULF2* exon 1 with last exon of *PRICKLE2*
NPEPPS-USP32	Intrachromosomal inversion	17q21.32-17q23.2	Fusion of *NPEPPS* exons 1–12 (out of 23) with *USP32* exons 2–4; 13-Mb inversion
ASTN2-PTPRG	Interchromosomal rearrangement	3p14.2-9q33.1	Fusion of *ASTN2* exons 1–10 (out of 22) with *PTPRG* exons 3–30
BCAS3-BCAS4	Interchromosomal rearrangement	17q23.2-20q13.13	*BCAS4* exon 1 fused to *BCAS3* exons 23–24; also found by Ruan et al. (2007)
BCAS3-RSBN1	Interchromosomal rearrangement	1p13.2-17q23.2	Fusion of *RSBN1* first exon with *BCAS3* exons 6–24
ASTN2-TBC1D16	Interchromosomal rearrangement	9q33.1-17q25.3	Fusion of *ASTN2* exons 1–15 with *TBC1D16* exons 2–12
BCAS4-PRKCBP1	Intrachromosomal inversion	20q13.12-20q13.13	Fusion of *BCAS4* exon 1 with *PRKCBP1* exons 5–22; 3.5-Mb inversion

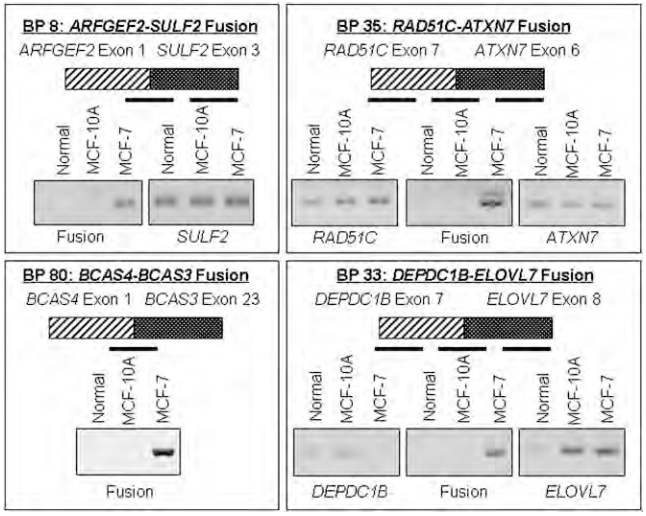

Figure 4. Confirmation of the presence of predicted processed chimeric mRNA transcripts in MCF-7 using RT-PCR.

fusion breakpoints and truncation breakpoints. The *PTPRG* breakpoints occur within the chromosome 3 breakpoint cluster and coincide within a known fragile site. Another example is the fusion of the *BMP7* promoter upstream of *ZNF217* breast cancer oncogene overexpressed in breast cancer (Collins et al. 2001) that we rediscovered but was also previously described Volik et al. (2003, 2006). The chromosome 20 rearrangement hotspot contains 37 breakpoints surrounding the *ZNF217* oncogene. Another extreme example of multiple rearrangements is the breast cancer amplified sequence 3 (*BCAS3*), occurring within the chromosome 17 rearrangement hotspot. There are seven breakpoints located within the intron–exon boundaries and an additional 19 nonfusion breakpoints surrounding the *BCAS3* gene region.

Rearrangements affect genes involved in homologous double-stranded break repair

We identified rearrangements in genes that code for members of protein complexes involved in double-stranded break repair (DSBR), raising the possibility that defects in DSBR genes may have contributed to genomic instability at certain stages of the evolution of the MCF-7 genome. One of the four MCF-7 gene fusions that produced a detectable predicted chimeric transcript is an interchromosomal fusion of *RAD51C* exons 1–7 to the neuronal-specific gene *ATXN7* exons 6–13. *RAD51C* is a paralog of *RAD51*, a gene central to DNA DSBR. RAD51C is an essential component of a complex reported to be involved in resolving holiday junctions (HJs) formed during DSBR (Liu et al. 2007) and as such is integral to the maintenance of genomic stability. The translocation we have identified eliminates the domain of *RAD51*C that binds other family member homologs such as RAD51D and Xrcc3 (Miller et al. 2004), possibly disrupting formation of the complex responsible for resolving HJs.

RAD51C is located at 17q23, a region of amplification that has been extensively studied in MCF-7 cells and breast cancer. One of the most studied oncogenes in breast cancer, *ErbB2,* is in close proximity to the 17q21.2 locus, which is amplified in a number of breast cancers (but not in MCF-7) but often independently of the 17q23 amplification. We examined *RAD51C* expression level in the microarray expression data set involving 50 breast cancer cell lines reported by Neve et al. (2006) and found that *RAD51C* levels are elevated in MCF-7, but much lower or absent in the majority of the other breast cancer cell lines.

We identified a translocation in another gene involved in DSBR, BRCA1-interacting protein-1 (*BRIP1,* also termed *BACH1*). BRIP1 was originally identified as a helicase-like protein that interacts directly with BRCA1 and contributes to its DNA repair function. BRIP1 binds to the BCRT repeat in BRCA1. The C terminus of BRIP1 is critical for its interaction with BRCA1, and a truncation mutant has been shown to block DSBR (Cantor et al. 2001; Yu et al. 2003; Lewis et al. 2005). Importantly, germline truncation mutations of *BRIP1* have been identified in familial breast cancer without mutations of *BRCA1/2*, and *BRIP1* truncations confer a twofold increased risk of developing breast cancer. We identified a translocation that results in the loss of the last three exons (exons 18–20); however, the fused DNA (3p14) downstream of *BRIP1* does not contain any exons or introns. The truncation at exon 17 of *BRIP1* would eliminate the C-terminal third of BRIP1 and eliminate binding to BRCA1. However, it is unclear at present whether the truncated mRNA would be stable as there is no transcription stop site or polyA tail.

Discussion

We have completed a sequence-level survey of rearrangements in a cancer genome. One major insight gained from this analysis is the presence of two types of breakpoints—clustered and dispersed, the latter being associated with LCRs. While we have not encountered previous reports of genome-wide association of LCRs with DSB breaks and chromosomal instability in tumors, the role of LCRs in promoting double-strand breaks through the replication fork stalling mechanism has recently been proposed in the context of genomic disorders (Lee et al. 2007).

A second major insight is that the two diverse types of breakpoints may have arisen during different stages of the evolution of the MCF-7 genome. Volik et al. (2006) hypothesized that 20q telomere loss initiated BFB cycles and a cascade of amplification resulting in small highly rearranged hotspots that colocalize DNA from different genomic regions. Our results show the same chromosomal rearrangement architecture, albeit at higher resolution and are consistent with the hypothesis that BFB cycles, possibly including extrachromosomal amplisomes, played an initial role in MCF-7 genome evolution. The chromosome 3 rearrangement hotspot encompasses the common fragile site *FRA3B*, prone to chromosomal instability, and a mediator of recurrent BFB amplification found in a variety of human tumors (Hellman et al. 2002). Recurrent breaks within common fragile sites propagated via BFB cycles amplify oncogenes and promote tumorgenesis (Huebner and Croce 2001; Hellman et al. 2002). Since both *RAD51C-ATXN7* fusion and *BRIP1* truncation belong to clusters possibly generated by the BFB mechanism, a possible effect is failure of the HR mechanism of DSBR and a consequent switch to NHEJ repair at stalled replication forks. A similar previously observed precedent is the switch from HR to NEHJ in *RAD54* homolog mutants (Sonoda et al. 2006). The switch to NHEJ at some point in the evolution of MCF-7 would have resulted in a mutator phenotype (Loeb 2001) and a pattern of extensive chromosomal rearrangements observed in MCF-7.

The switch to the rearrangement-creating NHEJ would have exposed the most breakage-prone sites—those containing LCRs—by converting simple replication-associated breaks into detectable

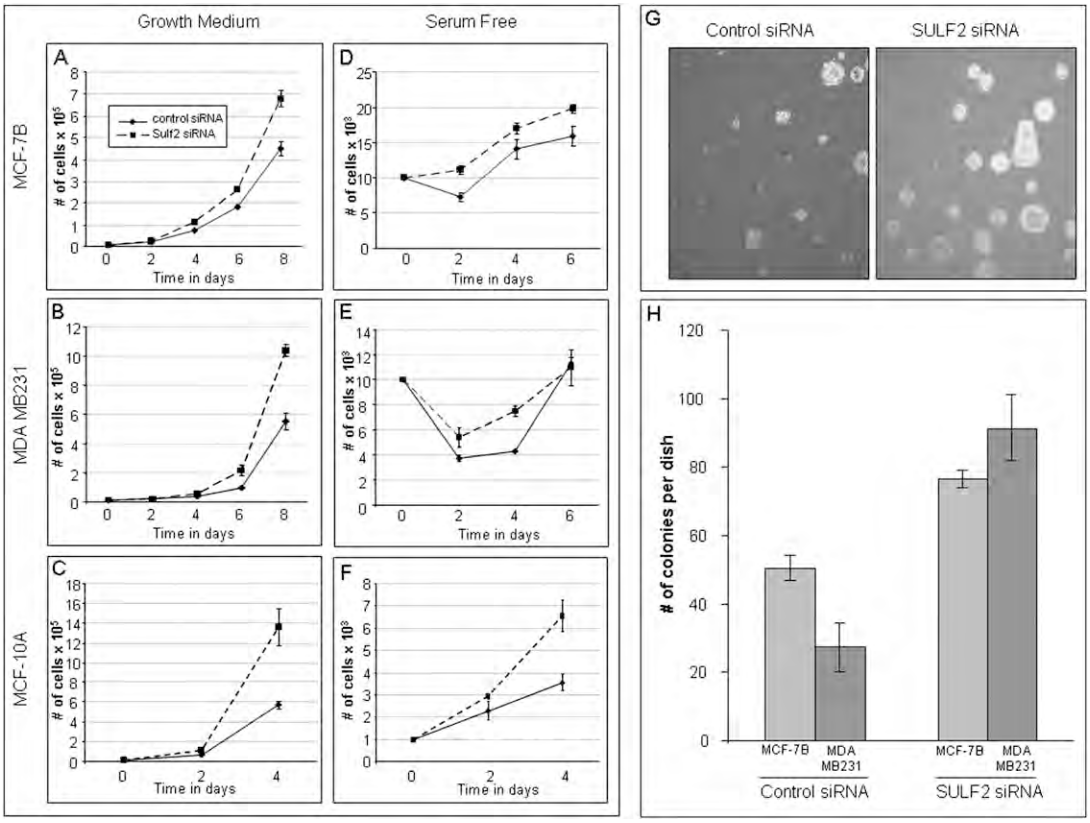

Figure 5. (*A–C*) Cells treated with *SULF2* siRNA have an enhanced proliferation compared with cells treated with control siRNA. MCF-7B (*A*; Mao et al. 2005), MDA MD231 (*B*), and MCF-10A (*C*) cells were transfected with 50 nM *SULF2* or control siRNA; 10^4 cells were plated in medium containing 10% FBS 48 h after transfection of the siRNA. Cells were counted on day 2, 4, 6, and 8. Experiments performed in triplicate; error bars show standard deviation. (*D–F*) Cells treated with *SULF2* siRNA have an enhanced survival compared with cells treated with control siRNA. MCF-7B (*D*; Mao et al. 2005), MDA MD231 (*E*), and MCF-10A (*F*) cells were transfected with 50 nM *SULF2* or control siRNA; 10^4 cells were plated in serum-free medium 48 h after transfection of the siRNA. Cells were counted on day 2, 4, and 6. Experiments performed in triplicate. Error bars, SD. (*G,H*) Treatment of MCF-7B and MDA MB231 cells with siRNA for *SULF2* increases the anchorage-independent growth capabilities. After treatment with siRNA, 10^4 cells were plated in 0.3% agar in growth medium, MCF-7B colony formation is shown in *G*. Plates were incubated for 21 d, and colonies were counted; bar chart results shown in *H*. Experiments performed in triplicate. Error bars, SD.

rearrangements. An analogy here exists between LCRs and DSB repair on one hand and microsatellites and mismatch repair on the other (Lengauer et al. 1998): By presenting challenges to DNA replication, LCRs and microsatellites, expose weaknesses in DSB repair and mismatch repair mechanisms, respectively. We should note that our extensive sequencing did not indicate increased mutability of MCF-7 at the base pair level, indicating highly functional mismatch repair.

The two-stage model also accounts for the typical curve indicating increase in genome complexity during the typical evolution of a breast cancer genome (Chin et al. 2004). While the BFB may account for the steep slope of rise in genomic complexity in MCF-7 during the stage of in situ carcinoma and telomere crisis, the subsequent instability mediated by the failure of the homology-based DSB repair mechanism resulting in breaks at LCR loci may account for the subsequent less steep slope that typically follows completion of the telomere crisis stage and accompanies metastasis. The two-stage model is also consistent with ongoing plasticity of the MCF-7 genome as evidenced by polyclonality and divergence of MCF-7 sublines (Jones et al. 2000; Nugoli et al. 2003).

The third insight is abundance of genes affected by rearrangements, and particularly of gene fusions, which exceeds current estimates of the abundance of gene fusions in breast cancer (Mitelman et al. 2007). Our unbiased screen of MCF-7 cell lines identified seventy nine genes involved in rearrangement events. Ten gene fusions were identified, nine novel and one previously reported by Ruan et al. (2007), and 77 other fusions involving genes and gene truncations.

The fourth insight is that at least a fraction of genes affected by fusions and truncations may in fact be tumor suppressors (e.g., *PTPRG*, *SULF2*) or may be responsible for genome stability (e.g., *RAD51C*, *BRIP1*). Both *BRIP1* and *RAD51C* fall within the cluster of breakpoints at 17q23 and are amplified in MCF-7 cells, indicating possible positive selection for the amplification. Such positive selection would be consistent with previously reported dominant-negative effects observed in genes responsible for genome stability (Milne and Weaver 1993).

The fifth insight is that chimeric transcripts can in fact be discovered by directly mapping rearrangements at the level of genomic DNA and then predicting specific chimeric transcripts. This opens the possibility of discovering recurrent, mechanistically and prognostically significant rearrangements by simply mapping a sufficient number of genomes and directly observing recurrent events.

In conclusion, this study validates the utility of mapping rearrangements in cancer genomes by providing mechanistically significant insights into cancer evolution and identifying genes likely involved in cancer progression. Building on the benchmarks developed in this study, next steps include technological and methodological improvements that will allow scale-up to whole genomes and to multiple cell lines and tumor samples at a more affordable cost, thus broadening applications in the research context and eventually in clinical settings.

Methods

Fosmid library preparation and end-sequencing of clone inserts

Fosmid libraries were prepared from each of the six 96-BAC pools indicated in Figure 1B using the Epicentre EpiFOS Fosmid Library Production Kit.

DNA sequencing

The ends of fosmid inserts were obtained using Sanger-based sequencing on an ABI 3730XL. Approximately 300,000 short (100-bp) reads were obtained from each of the three 192-BAC pools indicated in Figure 1B using the 454 Life Sciences (Roche) GS machine. Detailed sequencing statistics are included in the Supplemental Table 1. The sequencing reads are available for download from the public project pages at http://www.genboree.org.

Mapping reads onto the reference genome

Fosmid-end reads, 454 Life Sciences (Roche) shotgun reads, and BAC-end reads were mapped onto the reference human genome (March 2006 assembly, Build 36) using the BLAT program. BLAT parameters used for mapping are described in Supplementary Materials and coordinates are available through the Genboree site on the Breast Cancer project page at http://www.genboree.org.

PCR primer design pipeline

PCR primers were designed for amplifying breakpoint regions using repeat-masked human genome assembly (March 2006 assembly, Build 36) using a semi-automated primer design pipeline. Primer 3 primer design program was run to obtain a set of nested primers using two categories or parameters, "stringent" and "relaxed." Primer pairs in each category were scored, and the highest-scored primer pair was selected for initial round of PCR amplification. Priority was also given to the stringent category. In case of failure, additional lower-scoring primer pairs were employed. More details, including Primer 3 parameters, can be found in Supplemental materials.

PCR amplification of genomic DNA from cell lines

Breakpoint confirmation included PCR amplification of a pool of genomic DNA from six different sublines of MCF-7 cells (B, BK, C, D, L, and Neo). DNA isolated from immortalized but nontransformed mammary epithelial cells (MCF-10A) and normal human female DNA (Novagen) were used as negative controls. Genomic cell line DNA was isolated with the DNeasy kit (Qiagen). PCR bands were visualized on a 2% agarose gel.

Breakpoint clustering algorithm

Consecutive breakpoints that are closer than 1.1 Mbp in the reference genome assembly were connected. Runs of consecutive connected breakpoints with eight or more members are declared to constitute a cluster. Four clusters on chromosomes 1, 3, 17, and 20 indicated in Figure 3 were obtained in this fashion.

Identification of LCR regions

Each of the 157 MCF-7 breakpoints was examined for the presence of LCR. Intrachromosomal and interchromosomal LCRs were detected by applying a novel algorithmic method to the human genome assembly (March 2006 assembly, Build 36). The method involved self-comparison of the human genome using the Pash program (Kalafus et al. 2004) and an automated pipeline for segmentation, clustering, and parsing of LCRs based on sequence feature analysis. The LCRs detected by this method cover 6.15% of the whole genome in length, of which 18.7% are gene-containing regions. A detailed description of the algorithm is available in Supplemental materials.

Analysis of recurrent copy number changes in 157 somatic breakpoint loci

Copy number variation in the 157 somatic breakpoint loci identified in this study was examined. In order to identify recurrent copy number changes in breakpoint loci, array CGH data from 201 breast cancer cell lines and tumors (Chin et al. 2006; Neve et al. 2006; Shadeo and Lam 2006; Jonsson et al. 2007) were integrated. A locus was declared recurrently amplified if amplification was reported in more than 20% cases for the specific locus. Detailed results are compiled in a table where breakpoints are sorted by their level of recurrent copy number amplification (for details, see Supplemental materials and Supplemental Table 3).

Analysis of recurrent expression and copy number changes in 79 breakpoint-associated genes

Patterns of recurrent copy number and expression level variation were examined for 79 genes associated with the 157 somatic breakpoints identified in this study. Expression data from 50 breast cancer cell lines (Neve et al. 2006) were combined with copy number data from 201 breast cancer cell lines and tumors (Chin et al. 2006; Neve et al. 2006; Shadeo and Lam 2006; Jonsson et al. 2007). Detailed results are compiled in a table where genes are sorted by their level of recurrent alteration. (for details, see Supplemental Materials and Supplemental Table 2). Additionally, copy number data from an Affymetrix 100k SNP chip were used to identify breakpoint genes that also associate with regions of copy number alteration (see Supplemental Table 3).

Detection of predicted fusion transcripts by RT-PCR

mRNA from exponentially growing MCF-7 and MCF-10A cells were isolated with the RNeasy kit (Qiagen). To determine the presence of a fusion transcript, primers were designed across the fusion point on cDNA using Primer3. Control primers were designed on either side of the fusion. cDNA was generated by using gene specific primers. PCR amplification of the mRNA was restricted to 35 cycles. PCR bands were visualized on a 2% agarose gel, and verified by sequencing to confirm that the product contained mRNA from both genes involved.

Cell growth and soft-agar experiments

For the cell growth experiments, 10,000 cells were plated in triplicate in 24-well plates. The cells were grown in growth medium, containing 10% FBS, or in serum-free medium. Growth rate was

measured on days 0, 2, 4, and 6 with a Coulter Counter (Beckman Coulter).

Colony growth assays were performed as followed: 1 mL of solution of 0.5% noble agar in growth or serum-free medium was layered onto 30 × 10-mm tissue culture plates. A total of 1×10^4 cells was mixed with 1 mL of 0.3% agar solution prepared in a similar manner and layered on top of the 0.5% agar layer. Plates were incubated at 37°C in 5% CO_2 for 21 d. The experiment was performed in triplicate.

Knock-down of SULF2 using short interfering RNA (siRNA)

Transfections with *SULF2* and control nonspecific siRNA (Dharmacon) were carried out using 50 nM pooled siRNA duplexes and 4 μL of Dharmafect (Dharmacon) in six-well plates according to the manufacturer's protocol. After 48 h, the cells were prepared the respective assays.

Acknowledgments

We thank Andrew R. Jackson and Dr. Manuel Gonzalez-Garay for their computational support in providing the Genboree Discovery System, and Dr. Martin Krzywinski for providing the Circos circular genome visualization software. This project was funded by the NIH-NHGRI grant 1 R01 HG02583 and NIH-NCI grants R33 CA114151 and R21 CA128496 to A.M.

References

Bignell, G.R., Santarius, T., Pole, J.C., Butler, A.P., Perry, J., Pleasance, E., Greenman, C., Menzies, A., Taylor, S., Edkins, S., et al. 2007. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res.* **17:** 1296–1303.

Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C., et al. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40:** 722–729.

Cantor, S.B., Bell, D.W., Ganesan, S., Kass, E.M., Drapkin, R., Grossman, S., Wahrer, D.C., Sgroi, D.C., Lane, W.S., Haber, D.A., et al. 2001. BACH1, a novel helicase-like protein, interacts directly with BRCA1 and contributes to its DNA repair function. *Cell* **105:** 149–160.

Chin, K., de Solorzano, C.O., Knowles, D., Jones, A., Chou, W., Rodriguez, E.G., Kuo, W.L., Ljung, B.M., Chew, K., Myambo, K., et al. 2004. In situ analyses of genome instability in breast cancer. *Nat. Genet.* **36:** 984–988.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P.T., Roydasgupta, R., Kuo, W.L., Lapuk, A., Neve, R.M., Qian, Z., Ryder, T., et al. 2006. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10:** 529–541.

Collins, C., Volik, S., Kowbel, D., Ginzinger, D., Ylstra, B., Cloutier, T., Hawkins, T., Predki, P., Martin, C., Wernick, M., et al. 2001. Comprehensive genome sequence analysis of a breast cancer amplicon. *Genome Res.* **11:** 1034–1042.

Hahn, Y., Bera, T.K., Gehlhaus, K., Kirsch, I.R., Pastan, I.H., and Lee, B. 2004. Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases. *Proc. Natl. Acad. Sci.* **101:** 13257–13261.

Hellman, A., Zlotorynski, E., Scherer, S.W., Cheung, J., Vincent, J.B., Smith, D.I., Trakhtenbrot, L., and Kerem, B. 2002. A role for common fragile site induction in amplification of human oncogenes. *Cancer Cell* **1:** 89–97.

Hicks, J., Krasnitz, A., Lakshmi, B., Navin, N.E., Riggs, M., Leibu, E., Esposito, D., Alexander, J., Troge, J., Grubor, V., et al. 2006. Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* **16:** 1465–1479.

Huang, J., Wei, W., Zhang, J., Liu, G., Bignell, G.R., Stratton, M.R., Futreal, P.A., Wooster, R., Jones, K.W., and Shapero, M.H. 2004. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum. Genomics* **1:** 287–299.

Huebner, K. and Croce, C.M. 2001. FRA3B and other common fragile sites: The weakest links. *Nat. Rev. Cancer* **1:** 214–221.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Jones, C., Payne, J., Wells, D., Delhanty, J.D., Lakhani, S.R., and Kortenkamp, A. 2000. Comparative genomic hybridization reveals extensive variation among different MCF-7 cell stocks. *Cancer Genet. Cytogenet.* **117:** 153–158.

Jonsson, G., Staaf, J., Olsson, E., Heidenblad, M., Vallon-Christersson, J., Osoegawa, K., de Jong, P., Oredsson, S., Ringner, M., Hoglund, M., et al. 2007. High-resolution genomic profiles of breast cancer cell lines assessed by tiling BAC array comparative genomic hybridization. *Genes Chromosomes Cancer* **46:** 543–558.

Kalafus, K.J., Jackson, A.R., and Milosavljevic, A. 2004. Pash: Efficient genome-scale sequence anchoring by positional hashing. *Genome Res.* **14:** 672–678.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420–426.

Kytola, S., Rummukainen, J., Nordgren, A., Karhu, R., Farnebo, F., Isola, J., and Larsson, C. 2000. Chromosomal alterations in 15 breast cancer cell lines by comparative genomic hybridization and spectral karyotyping. *Genes Chromosomes Cancer* **28:** 308–317.

Lee, J.A., Carvalho, C.M., and Lupski, J.R. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131:** 1235–1247.

Lengauer, C., Kinzler, K.W., and Vogelstein, B. 1998. Genetic instabilities in human cancers. *Nature* **396:** 643–649.

Lewis, A.G., Flanagan, J., Marsh, A., Pupo, G.M., Mann, G., Spurdle, A.B., Lindeman, G.J., Visvader, J.E., Brown, M.A., and Chenevix-Trench, G. 2005. Mutation analysis of FANCD2, BRIP1/BACH1, LMO4 and SFN in familial breast cancer. *Breast Cancer Res.* **7:** R1005–R1016.

Liu, Y.T. and Carson, D.A. 2007. A novel approach for determining cancer genomic breakpoints in the presence of normal DNA. *PLoS One* **2:** e380. doi: 10.1371/journal.pone.0000380.

Liu, Y., Tarsounas, M., O'Regan, P., and West, S.C. 2007. Role of RAD51C and XRCC3 in genetic recombination and DNA repair. *J. Biol. Chem.* **282:** 1973–1979.

Loeb, L.A. 2001. A mutator phenotype in cancer. *Cancer Res.* **61:** 3230–3239.

Mao, J.H., Li, J., Jiang, T., Li, Q., Wu, D., Perez-Losada, J., DelRosario, R., Peterson, L., Balmain, A., and Cai, W.W. 2005. Genomic instability in radiation-induced mouse lymphoma from p53 heterozygous mice. *Oncogene* **24:** 7924–7934.

Miller, K.A., Sawicka, D., Barsky, D., and Albala, J.S. 2004. Domain mapping of the Rad51 paralog protein complexes. *Nucleic Acids Res.* **32:** 169–178.

Milne, G.T. and Weaver, D.T. 1993. Dominant negative alleles of RAD52 reveal a DNA repair/recombination complex including Rad51 and Rad52. *Genes & Dev.* **7:** 1755–1765.

Mitelman, F., Johansson, B., and Mertens, F. 2007. The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* **7:** 233–245.

Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.P., Tong, F., et al. 2006. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10:** 515–527.

Nugoli, M., Chuchana, P., Vendrell, J., Orsetti, B., Ursule, L., Nguyen, C., Birnbaum, D., Douzery, E.J., Cohen, P., and Theillet, C. 2003. Genetic variability in MCF-7 sublines: Evidence of rapid genomic and RNA expression profile modifications. *BMC Cancer* **3:** 13. doi: 10.1186/1471-2407-3-13.

Raphael, B.J., Volik, S., Yu, P., Wu, C., Huang, G., Linardopoulou, E.V., Trask, B.J., Waldman, F., Costello, J., Pienta, K.J., et al. 2008. A sequence-based survey of the complex structural organization of tumor genomes. *Genome Biol.* **9:** R59. doi: 10.1186/gb-2008-9-3-r59.

Ruan, Y., Ooi, H.S., Choo, S.W., Chiu, K.P., Zhao, X.D., Srinivasan, K.G., Yao, F., Choo, C.Y., Liu, J., Ariyaratne, P., et al. 2007. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res.* **17:** 828–838.

Rummukainen, J., Kytola, S., Karhu, R., Farnebo, F., Larsson, C., and Isola, J.J. 2001. Aberrations of chromosome 8 in 16 breast cancer cell lines by comparative genomic hybridization, fluorescence in situ hybridization, and spectral karyotyping. *Cancer Genet. Cytogenet.* **126:** 1–7.

Shadeo, A. and Lam, W.L. 2006. Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Res.* **8:** R9.

Sonoda, E., Hochegger, H., Saberi, A., Taniguchi, Y., and Takeda, S. 2006. Differential usage of non-homologous end-joining and homologous recombination in double strand break repair. *DNA Repair (Amst.)* **5:** 1021–1029.

Soule, H.D., Vazguez, J., Long, A., Albert, S., and Brennan, M. 1973. A human cell line from a pleural effusion derived from a breast carcinoma. *J. Natl. Cancer Inst.* **51:** 1409–1416.

Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., et al. 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310:** 644–648.

van Doorn, R., Zoutman, W.H., Dijkman, R., de Menezes, R.X., Commandeur, S., Mulder, A.A., van der Velden, P.A., Vermeer, M.H., Willemze, R., Yan, P.S., et al. 2005. Epigenetic profiling of cutaneous T-cell lymphoma: Promoter hypermethylation of multiple tumor suppressor genes including BCL7a, PTPRG, and p73. *J. Clin. Oncol.* **23:** 3886–3896.

Volik, S., Zhao, S., Chin, K., Brebner, J.H., Herndon, D.R., Tao, Q., Kowbel, D., Huang, G., Lapuk, A., Kuo, W.L., et al. 2003. End-sequence profiling: Sequence-based analysis of aberrant genomes. *Proc. Natl. Acad. Sci.* **100:** 7696–7701.

Volik, S., Raphael, B.J., Huang, G., Stratton, M.R., Bignel, G., Murnane, J., Brebner, J.H., Bajsarowicz, K., Paris, P.L., Tao, Q., et al. 2006. Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res.* **16:** 396–404.

Wary, K.K., Lou, Z., Buchberg, A.M., Siracusa, L.D., Druck, T., LaForgia, S., and Huebner, K. 1993. A homozygous deletion within the carbonic anhydrase-like domain of the Ptprg gene in murine L-cells. *Cancer Res.* **53:** 1498–1502.

Yu, X., Chini, C.C., He, M., Mer, G., and Chen, J. 2003. The BRCT domain is a phospho-protein binding domain. *Science* **302:** 639–642.

Received April 29, 2008; accepted in revised form November 19, 2008.

ELSEVIER

Cancer Genetics 204 (2011) 447–457

Cancer Genetics

Long-range massively parallel mate pair sequencing detects distinct mutations and similar patterns of structural mutability in two breast cancer cell lines

Oliver A. Hampton [a,b,c,*], Christopher A. Miller [a,b,d], Maxim Koriabine [e], Jian Li [a,b], Petra Den Hollander [f,g], Lucia Carbone [e,h], Mikhail Nefedov [e], Boudewijn F.H. Ten Hallers [e], Adrian V. Lee [f,i], Pieter J. De Jong [e], Aleksandar Milosavljevic [a,b]

[a] *Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX;* [b] *Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX;* [c] *Bionanomatrix Inc., Philadelphia, PA;* [d] *Washington University School of Medicine, The Genome Center, St. Louis, MO;* [e] *BACPAC Resources, Children's Hospital of Oakland Research Institute, Oakland, CA;* [f] *Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX;* [g] *Department of Clinical Cancer Prevention, The University of Texas MD Anderson Cancer Center, Houston, TX;* [h] *Department of Behavioral Neuroscience, Oregon Health and Science University, Portland, OR;* [i] *Magee Womens Research Institute, University of Pittsburgh Cancer Institute, Pittsburgh, PA, USA*

Cancer genomes frequently undergo genomic instability resulting in accumulation of chromosomal rearrangement. To date, one of the main challenges has been to confidently and accurately identify these rearrangements by using short-read massively parallel sequencing. We were able to improve cancer rearrangement detection by combining two distinct massively parallel sequencing strategies: fosmid-sized (36 kb on average) and standard 5 kb mate pair libraries. We applied this combined strategy to map rearrangements in two breast cancer cell lines, MCF7 and HCC1954. We detected and validated a total of 91 somatic rearrangements in MCF7 and 25 in HCC1954, including genomic alterations corresponding to previously reported transcript aberrations in these two cell lines. Each of the genomes contains two types of breakpoints: clustered and dispersed. In both cell lines, the dispersed breakpoints show enrichment for low copy repeats, while the clustered breakpoints associate with high copy number amplifications. Comparing the two genomes, we observed highly similar structural mutational spectra affecting different sets of genes, pointing to similar histories of genomic instability against the background of very different gene network perturbations.

Keywords Copy number variation, fosmid diTag, gene fusion, genomic instability, massively parallel sequencing

End sequence profiling of clonal libraries have been used extensively to discover structural variation in both normal and cancer genomes (1–5). Recently, the adoption of massively parallel sequencing has supplemented structural variation detection by identifying rearrangements at fine scale resolution for both normal and cancer genomes (6–14). These massively parallel sequencing studies have significantly added to the catalog of genomic rearrangements, but the limited insert sizes between paired ends have provided less power than larger insert clones to map across duplications and repeat-rich regions in the genome, thereby missing a large fraction of variation (2,15). Massively parallel mate pair sequencing is also hindered by high false-positive rearrangement discovery rates, requiring additional breakpoint

Received April 22, 2011; received in revised form July 7, 2011; accepted July 21, 2011.

* Corresponding author.
E-mail address: ohampton@bionanomatrix.com

2210-7762/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cancergen.2011.07.009

validation. Commonly used techniques for breakpoint validation include optical mapping based on restriction enzyme maps or incorporated fluorochrome-labeled nucleotide (2,16), hybridization of fluorescent probes that span rearrangements (5,17), and polymerase chain reaction amplification across aberrant fusions followed by Sanger-based sequencing of the breakpoint amplicons (1,7,11,12,14). Although these validation techniques offer proof of genomic rearrangement, none are currently amenable to high-throughput workflows.

We supplement the limited insert size of standard massively parallel mate pair sequencing by incorporating fosmid-sized insert libraries, thereby providing additional validation of detected rearrangements. Our fosmid-sized mate pair libraries, called fosmid diTags, leverage the affordable costs and high-throughput capacity of massively parallel sequencing while providing clone-sized inserts able to span long, repetitive sequence elements. Fosmid diTags are well suited for rearrangement detection either in standalone or complementary fashion with other mate pair libraries; fosmid diTags are also advantageous for de novo genome assembly where larger insert size facilitates greater continuity (18). Fosmid diTags are an extension of paired end tag methods (14,19–21), where short paired tags from the ends of DNA fragments are enzymatically extracted and covalently linked as diTag constructs for high-throughput sequencing. Fosmid diTag workflow details are provided in Supplemental Figures 1 and 2.

Materials and methods

Sequencing library preparation

Paired end sequencing methods exploit the fact that structural abnormalities consist of two chromosomal segments that are in a relative position and orientation, or at a relative distance that is not consistent with the reference genome assembly. Construction of paired end sequencing libraries that adequately cover the genome of interest allows for comprehensive identification of structural abnormalities.

A total of 1.55 million MCF7 (ATCC [American Type Culture Collection, Manassas, VA] HTB-22) and 1.50 million HCC1954 (ATCC CRL-2338) fosmids were cloned by using the novel pFosDT1.2 vector (derived from the Epicentre pCC1FOS plasmid). The pFosDT1.2 vector contains two *Eco*P15I restriction sites that flank the site of insertion. *Eco*P15I, a type III restriction endonuclease, cuts 25 and 27 bp downstream of its recognition site producing a 2 bp 5′ overhang and requires two separated and inversely oriented recognition sites in supercoiled DNA for native cleavage. Addition of sinefungin in the *Eco*P15I digest reaction facilitates cleavage at all recognition sites independent of DNA topology (22). Starting with 10 μg of purified pooled fosmid DNA from each breast cancer cell line, two independent long-range clonal insert fosmid diTag massively parallel sequencing libraries were produced. For each fosmid library, 26 bp end tags from the insert termini were isolated and concatenated (Supplemental Figure 2).

Illumina mate pair whole-genome shotgun libraries, of insert sizes ranging from 4 to 6 kb, were additionally constructed with 10 μg of genomic DNA from each of the MCF7 (ATCC HTB-22) and HCC1954 (ATCC CRL-2338) cell lines. Mate pair libraries were prepared according to the manufacturer's instructions (Illumina PE-112-1002). Two separate MCF7 mate pair libraries with 4 kb and 6 kb inserts were constructed, and a single HCC1954 mate pair library with 5 kb inserts was constructed.

The fosmid diTag and Illumina mate pair libraries were sequenced on an Illumina Genome Analyzer II massively parallel sequencing system following the manufacturer's instructions. Raw sequence data for the fosmid diTag and standard Illumina mate pair libraries are available online (http://www.genboree.org/breastCellLineReads/).

Mapping to reference genome

Novocraft V2.05.02 was used to align quality-filtered paired end reads to the reference human genome (March 2006 assembly, National Center for Biotechnology Information [NCBI] build 36.1, University of California–Santa Cruz [UCSC] build HG18). Novoalign parameters used for mapping are described in the Supplementary Materials, and mapping coordinates are available for viewing and download through the Genboree open-hosting genome browser (http://www.genboree.org/).

Structural rearrangement calling

Fosmid diTags and Illumina mate pair sequences that align discordantly were used to call putative structural rearrangements. The combined fosmid diTag and standard Illumina mate pair structural rearrangements that validated as cancer-specific somatic mutations are available in Supplemental Table 1.

Determining structural variants from Illumina mate pair and fosmid diTag sequences is complicated by two factors: the contamination of inward-facing reads and the formation of chimeric clones, respectively. Inward-facing reads are paired end sequences from a contiguous piece of DNA sized equal to the final sequencing library length, approximately 400 bp. Formation of chimeric clones during the fosmid diTag procedure introduce false information about the distance and orientation between two reads, complicating structural variant calling.

False-positive breakpoints called by inward-facing reads were removed before reporting structural variants by filtering the discordant read clusters. Inward-facing read clusters were filtered on the basis of size and their inward-facing read orientation. Because inward-facing read clusters are limited by the final sequencing library length, they are easily identified and may be removed from further analysis (see Supplementary Materials for filter parameters). However, inward-facing read clusters that span truly positive rearrangements will also be removed, thereby introducing detection voids. To overcome fosmid diTag chimera noise, discordant tags supporting the same structural variant were clustered. Clusters are formed if there are at least two uniquely mapping paired end signatures with corroborating genomic positions, sizes, and read orientations. Such a strategy is called standard clustering, and it is commonly used (6,8,10,23–25).

Breakpoint spanning polymerase chain reaction primer design pipeline

Polymerase chain reaction (PCR) primers were designed for amplification across aberrant fusions by using the human reference genome (March 2006 assembly, NCBI build 36.1, UCSC build HG18). The Primer3 primer design algorithm was used to obtain a set of nested primers with two categories of parameters, stringent and relaxed. Primer pairs in each category are scored, and the highest scoring primer pair is selected for PCR assay validation. Priority is given to the stringent category using the repeat-masked human reference genome. In cases of PCR amplification failure, additional lower scoring primer pairs were utilized. More details, including Primer3 parameters, can be found in the Supplementary Materials; the automated primer design pipeline code is available for download (http://github.com/oliverhampton/Breakpoint-Primer-Design).

Copy number variation calling

Uniquely mapping reads were used as input for the readDepth R package (26), which calls copy number alterations by evaluating depth of sequence coverage. The package's default parameters were used including an overdispersion value of 3 and a false discovery rate of 0.01. The readDepth package also provided a breakpoint refinement tool that allowed us to adjust copy number segment ends to matched breakpoint positions.

Breakpoint clustering algorithm

Combined fosmid-sized and 5 kb breakpoints that map within 2 Mb in MCF7 and 5 Mb in HCC1954 were clustered. Chromosome segment annotations were retained if five or more breakpoints in MCF7 or two or more breakpoints in HCC1954 were contained within the cluster. For each set of MCF7 and HCC1954 breakpoint clusters, the cluster containing the highest number of breakpoints served as a seed for a connected graph (or clique) where the chromosome segments are nodes and spanning breakpoints are edges. In this manner, cliques of four breakpoint clusters in MCF7 on chromosomes 1, 3, 17, and 20, and five breakpoint clusters in HCC1954 on chromosomes 5, 8, and 11 were constructed.

Identification of low copy repeat regions

Each of the fosmid diTag and Illumina mate pair breakpoints from MCF7 and HCC1954 breast cancer cell lines were examined for the presence of low copy repeats (LCRs). Intra- and interchromosomal homologous LCRs were detected by applying a novel algorithmic method to the human reference genome sequence (March 2006 assembly, NCBI build 36.1, UCSC build HG18). The method achieved higher sensitivity than previously applied methods (27) by using *k*-mer frequency sequence information to detect, parse, and cluster LCRs without removing high copy number repetitive elements (repeat masking). The LCRs detected by this method covered 6% of the whole genome in length, of which 19% were gene-containing regions. A detailed description of the algorithm is available in the Supplementary Materials.

PCR amplification of genomic DNA from cell lines

Breast cancer cell line breakpoint confirmation used PCR amplification of MCF7 (ATCC HTB-22) and a pool of negative control genomic DNA isolated from human female (Novagen 70605-3) and from two different cell lines, MCF10A (ATCC CRL-10317) and HCC1599-BL (ATCC CRL-2332); and PCR amplification of genomic DNA from HCC1954 (ATCC CRL-2338) and negative control HCC1954-BL (ATCC CRL-2339) cell lines. Genomic cell line DNA was isolated with the DNAeasy kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. PCR bands were visualized on a 2% agarose gel.

Results

Combining fosmid diTag and 5 kb mate pair sequencing libraries increases specificity to detect chromosomal rearrangements

The Illumina standard mate pair libraries, with an average 5 kb insert size, generated 2.9 and 1.9 Gb of sequence data for MCF7 and HCC1954, respectively. Upon mapping to the reference genome, the relatively short distance between the paired ends was compatible for PCR primer design across aberrant fusions, and the density of mapped reads allowed for the measurement of segment copy number. The fosmid diTag libraries generated 93.3 and 56.9 Mb of sequence data for MCF7 and HCC1954, respectively. Because of the larger insert size, mapping of fosmid diTags provided nearly identical percent insert coverage of the reference genome (81% mean $\pm$ 7% SD), as observed from the Illumina mate pair libraries.

Rearrangements were reported where at least two independent pairs of ends showed discrepancy by their predicted size and/or orientation. Discordant mate pairs and diTags were reported when the distance between the mapped ends was in excess of 2 standard deviations from the insert mean. Discordant ends were clustered on the basis of mapping position and orientation discrepancies, thereby refining the position of the detected breakpoint. From the Illumina 5 kb mate pair libraries, we could identify 23,555 putative rearrangements in MCF7 and 3,824 in HCC1954. Breakpoint spanning PCR primer designs were able to be created for 23% of these rearrangements in MCF7 and 61% in HCC1954. From the fosmid diTag libraries, we identified 713 putative rearrangements in MCF7 and 345 in HCC1954; because of the much longer fosmid diTag insert size and relatively low fold coverage, standard and long-range PCR primer designs were incompatible.

The high percentage of failed PCR primer designs from the MCF7 mate pair data was due to the increased prevalence of repetitive sequence elements surrounding aberrant fusions. Closer inspection of the PCR primer design failure sites revealed overlap with repeat-masked sections of the human genome and disproportionate calling of small indels (2–4 kb) at a rate 10 times more than expected. We speculate that MCF7 has unique defects in its DNA repair

pathways, which explains the imbalance of mutations between the two cell lines. We have previously shown *RAD51C* to be mutated in MCF7 (28); such a mutation could affect the Holliday junction (HJ) (29) resolution machinery, causing misrecognition of HJs, cruciforms, and other homology-driven secondary structures leading to double-strand breaks and accumulation of such indels (30).

Breakpoint spanning primers from the Illumina mate pairs were applied to their respective breast cancer cell line genomes and normal controls. In most cases, the PCR assay failed to produce an amplification product, indicating a high rate of false-positive rearrangement detection. In the cases where a breakpoint amplicon was produced, the majority identified normal structural polymorphisms—only a small percentage identified breast cancer-specific somatic mutation (Figure 1B). Interestingly, combining fosmid diTag and Illumina mate pair data, and selecting rearrangements detected by both methods showed a threefold enrichment for cancer-specific somatic mutation and a twofold reduction in false-positive detection when compared to the Illumina mate pair libraries alone. Combining fosmid-sized and 5 kb mate pairs provides cross-validation to rearrangement detection; moreover, the incorporation of longer fosmid-sized inserts increases specificity to detect breast cancer-specific somatic mutation and decreases the reporting of false-positive rearrangements when compared to the shorter 5 kb inserts alone. Combining fosmid diTag and 5 kb mate pair libraries, we identified 309 chromosomal rearrangements in MCF7 and 72 in HCC1954, and designed breakpoint spanning PCR primers for approximately 90% of them (Figure 1A). Although it is desirable to increase the specificity of chromosomal rearrangement detection, it must be noted that a corresponding loss of sensitivity is associated with this improvement.

Corresponding genomic DNA fusions exist for upward of half of the gene fusions and truncations previously detected by transcript mapping

Chimeric gene transcripts have been previously identified in MCF7 (31,32) and HCC1954 (33) by transcript mapping. Transcript mapping is analogous to targeted paired end sequencing; however, instead of investigating aberrant genomic fusions, chimeric mRNA transcripts are queried. Transcript mapping delivers a gene-centric view of rearrangements that encompass posttranscriptional modifications, but can't detect genomic rearrangements outside of gene coding regions. We therefore sought to comprehensively identify rearrangement events at the genomic DNA level that may have caused chimeric or truncated mRNA transcripts.

In MCF7, we identified 10 of 19 and 9 of 30 genomic rearrangements correlated with corresponding chimeric mRNA transcripts reported by Maher et al. (9,31) and Inaki et al. (32), respectively. These genomic lesions involve oncogenes (*TMEM49*), tumor suppressors (*SULF2, PTPRG*), constituents of DNA double-strand break repair (*RAD51C, BRIP1*), and other genes related to cell cycle, growth, and survival (*RPS6KB1, ELOVL7, ABCA5*) (Table 1).

In HCC1954, we identified 3 of 7 genomic rearrangements resulting in chimeric or truncated gene transcripts reported by Zhao et al. (33). These three gene truncations (*EIF3E, NSD1, PVT1*) are implicated in differing aspects of breast and ovarian cancers, and acute myeloid leukemia pathophysiologies (Table 1). In addition, we discovered a novel genomic rearrangement of *UIMC1* (*RAP80*), a DNA double-stranded break repair accessory protein and suspected tumor suppressor, resulting in the loss of its last 5 exons (exons 11−15), which would eliminate its DNA recognition and binding abilities. The fused DNA (8q24.21) downstream of the *UIMC1* breakpoint does not contain any exons or introns, and it remains unclear whether the truncated mRNA would be stable as there is no transcription stop site or polyA tail.

High-level amplifications of distinct driver oncogenes in MCF7 and HCC1954 are detected from mapped read density

The luminal-type MCF7 and *ERBB2*-overexpressing HCC1954 breast cancer cell lines are both highly amplified and display complex structural mutability phenotypes; exhibiting distinct profiles of genome structural rearrangement and copy number variation. We integrated read density and breakpoint information from mapped fosmid-sized and 5 kb mate pair libraries to accurately identify copy number variation by the readDepth R package (26). Figures 1A and 2 show visualized copy number counts.

For comparison, we obtained data for both breast cancer cell lines run on Affymetrix 100K single nucleotide polymorphism (SNP) chips segmented with the Gain and Loss Analysis of DNA (GLAD) algorithm, microarray data available in the NCBI GEO database (accession GSE13696) (http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE13696) (34). Even with approximately onefold sequence coverage, our results provided higher resolution than the Affymetrix arrays and allowed for copy number variation calls to be made in many regions where no array probes exist. A look at gross features showed good concordance between the two platforms, including detection of previously described high-level amplifications on MCF7 (28); and on HCC1954 (12) (highlighted regions in Figure 2B). Notably, our sequence-based approach provided higher dynamic range and revealed multiple regions in both cell lines that have been copied 50 to 100 times. As a result of saturation effects and lower resolution, these regions are called with far lower copy number on the Affymetrix arrays. There are a small number of aberrations, including regions on chromosomes 2 and 9 in the MCF7 genome, which we believe to be biological differences between different passages and/or sublines of MCF7. Most other discordant events are likely attributable to increased coverage, resolution, and dynamic range from the sequence-based assays.

In MCF7, a 20 kb segment on cytoband 20q13.31 showed the highest level of amplification with a copy number count of 70. This region encompasses the *BMP7* gene, a member of the transforming growth factor-beta superfamily, and corresponds to the fusion of the *BMP7* promoter upstream of the *ZNF217* oncogene, which is overexpressed in breast cancer (35). ZNF217 can attenuate apoptotic signals resulting from telomere dysfunction and

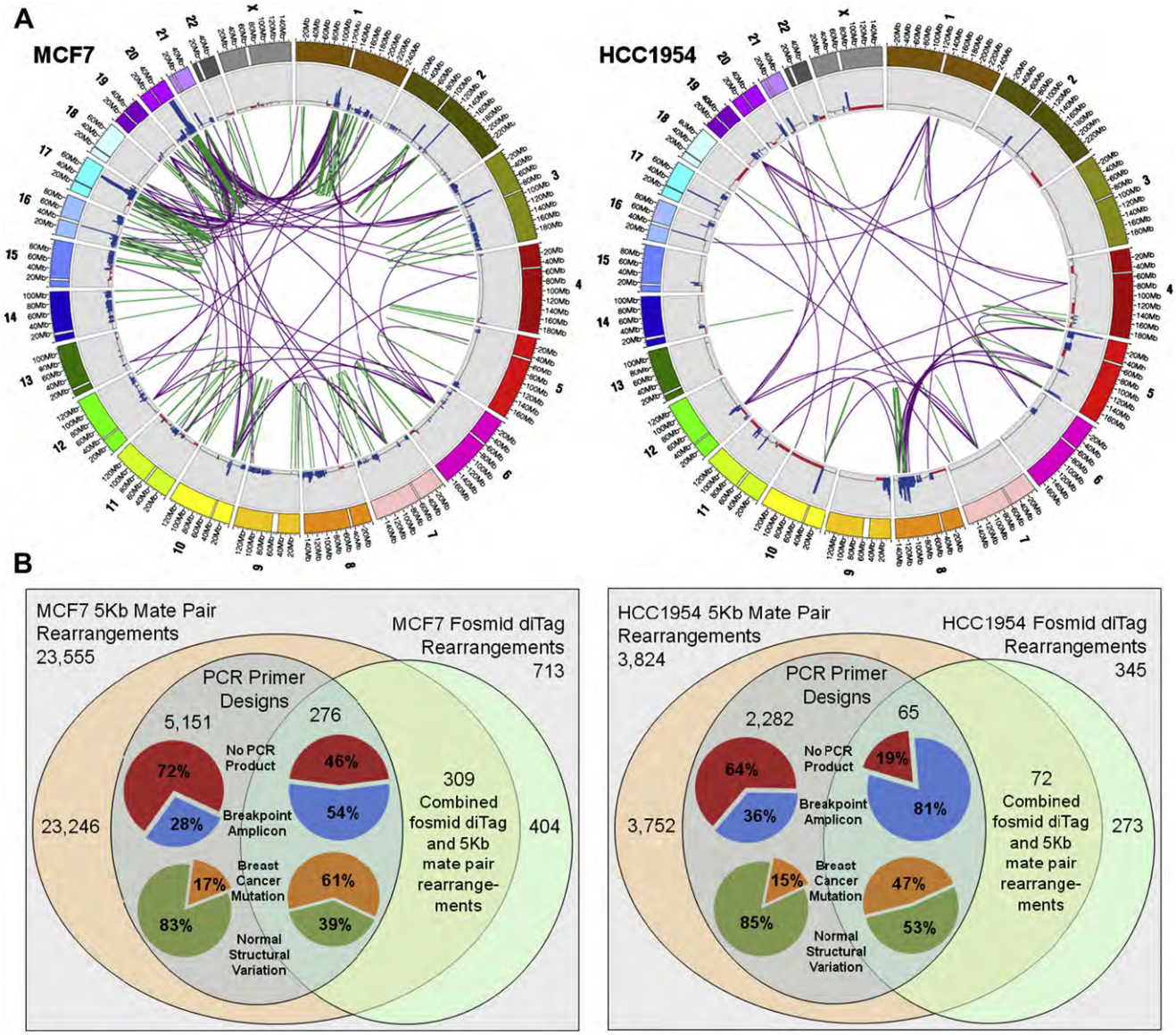

Figure 1 (A) Circular visualizations of the MCF7 and HCC1954 genomes obtained by Circos (52) software. Chromosomes are individually colored with centromeres in white. Copy number variation is plotted with gains in blue and losses in red. The colored rearrangements depict breast cancer-specific somatic mutations from the combined fosmid-sized and 5 kb mate pair libraries. Green lines denote intrachromosomal and purple lines denote interchromosomal rearrangements. (B) Venn diagrams comparing the numbers of fosmid-sized and 5 kb mate pair rearrangements, PCR primer designs, PCR assays that produced breakpoint amplicon vs. no amplification product, and rearrangements that are validated as breast cancer-specific mutation vs. normal structural variation in the MCF7 and HCC1954 genomes.

may promote neoplastic transformation during later stages of malignancy (36).

In HCC1954, a 51 kb segment on cytoband 17q12 showed the highest level of amplification with a copy number count of 117. This region encompasses *HER2/neu* (also known as *ERBB-2*), which is known to be overexpressed in this cell line. *HER2* overexpression in breast cancer is associated with an aggressive tumor phenotype, increased disease recurrence, and overall worse prognosis. *HER2* overexpression serves not only as a prognostic marker, but also as a drug target for the monoclonal antibody trastuzumab. Also of interest, a 59 kb segment on cytoband 11q13.2, encompassing the gene *CCND1*, showed ninefold amplification. The *CCND1* gene, a key cell-cycle regulator, is often overexpressed in breast cancer patients, and correlates with shorter relapse-free survival times (37).

MCF7 and HCC1954 exhibit defects in the homologous double-strand break repair pathway

In the MCF7 and HCC1954 breast cancer cell lines, we identified rearrangements in genes that code for members of protein complexes involved in DNA double-stranded break repair (DSBR), raising the possibility that distinct defects in

Table 1 Validated chimeric protein fusions and gene truncations in MCF7 and HCC1954 breast cancer cell lines

Cell line	Genomic changes	Chromosome locations	Genes affected	Effect on coding	Somatic changes reported in cancers	Validation method
MCF7	Interchromosomal translocation	t(17;20)(q23.2;q13.13)	*BCAS3, BCAS4*	Chimeric protein	Fusion is recurrently present in MCF7 breast and HCT116 colon cancer cell lines.	cDNA, genomic and published (9,20,28,32,53)
MCF7	Interchromosomal translocation	t(3;17)(p14.1;q22)	*ATXN7, RAD51C*	Chimeric protein	Decreased expression of *RAD51C* found in the majority of breast cancer cell lines.	FISH, cDNA, genomic and published (28,32)
MCF7	Intrachromosomal inversion	t(20;20)(q13.13;q13.13)	*SULF2, ARFGEF2*	Chimeric protein	SULF2 is a known tumor suppressor. *SULF2* siRNA silencing is tumorigenic *in vivo*.	cDNA, genomic and published (9,28,32)
MCF7	Interchromosomal translocation	t(3;20)(p14.1;q13.13)	*SULF2, PRICKLE2*	Chimeric protein	See above re: *SULF2* function and phenotype	Genomic and published (9,28,32)
MCF7	Intrachromosomal indel	t(19;19)(p13.11;p13.11)	*MYO9B, FCHO1*	Chimeric protein	*MYO9B* mutations associate with different inflammatory or autoimmune diseases.	Genomic and published (9,32)
MCF7	Intrachromosomal indel	t(17;17)(q22;q23.1)	BC017255*, TMEM49*	Chimeric protein	High-level amplification at *TMEM49* induces high expression of miR-21, which targets *PTEN* and results in an aggressive breast cancer phenotype (54).	Genomic and published (9)
MCF7	Intrachromosomal inversion	t(17;17)(q23.1;q23.1)	*RPS6KB1, TMEM49*	Chimeric protein	See above *TMEM49* function and phenotype. *RPS6KB1* is amplified and overexpressed in 10−30% of primary breast cancers and cell lines. RPS6KB1 is regulated by the *MTOR* pathway, which regulates cell cycle, growth, and survival (55).	Genomic and published (9,32)
MCF7	Intrachromosomal inversion	t(5;5)(q12.1;q12.1)	*DEPDC1B, ELOVL7*	Chimeric protein	*ELOVL7* is over expressed in bladder, breast, colorectal, esophageal, gastric, and prostate cancers. High-fat diet promotes growth of *in vivo* tumors of *ELOVL7*-expressed prostate cancer (56).	cDNA, genomic and published (9,28,32)

(*continued on next page*)

Table 1 (*Continued*)

Cell line	Genomic changes	Chromosome locations	Genes affected	Effect on coding	Somatic changes reported in cancers	Validation method
MCF7	Interchromosomal translocation	t(3;17)(p14.2;q22)	*TEX14, PTPRG*	Chimeric protein	*PTPRG* is a known tumor suppressor in kidney, lung and breast cancers. PTPRG has been shown to inhibit MCF7 anchorage-independent growth and reduce estrogenic response cell proliferation (57).	Genomic and published (9,32)
MCF7	Interchromosomal translocation	t(17;20)(q24.3;q13.32)	*ABCA5, PPP4R1L*	Chimeric protein	Induction of *ABCA5* correlates with differentiation state of human colon tumor.	Genomic and published (9)
MCF7	Intrachromosomal inversion	t(X;X)(p22.2;p22.2)	*CXorf15, SYAP1*	Chimeric protein	No known cancer phenotype.	Genomic and published (9,32)
MCF7	Interchromosomal translocation	t(3;17)(p14.1;q23.2)	*BRIP1*	Truncation	*BRIP1* truncations confer a twofold increased risk of developing breast cancer. Truncation mutants block double stranded break repair.	Genomic and published (9)
HCC1954	Interchromosomal translocation	t(5;8)(q23.1;q13.13)	*EIF3E*	Truncation	Truncation is tumorigenic *in vivo*. Decreased expression found in one-third of all human breast carcinomas.	cDNA, genomic and published (12,33)
HCC1954	Interchromosomal translocation	t(5;8)(q35.3;q24.21)	*NSD1*	Truncation	Fusion protein in acute myeloid leukemia.	FISH, cDNA, genomic and published (33)
HCC1954	Interchromosomal translocation	t(5;8)(p15.33;q24.21)	*CLPTM1L, PVT1*	Truncation	Amplification of *PVT1* linked to pathophysiology of ovarian and breast cancers.	Genomic and published (33)
HCC1954	Interchromosomal translocation	t(5;8)(p15.35.2;q24.21)	*UIMC1* or *RAP80*	Truncation	Recurrent *RAP80* missense mutations identified in breast cancer patients (45,46).	Genomic

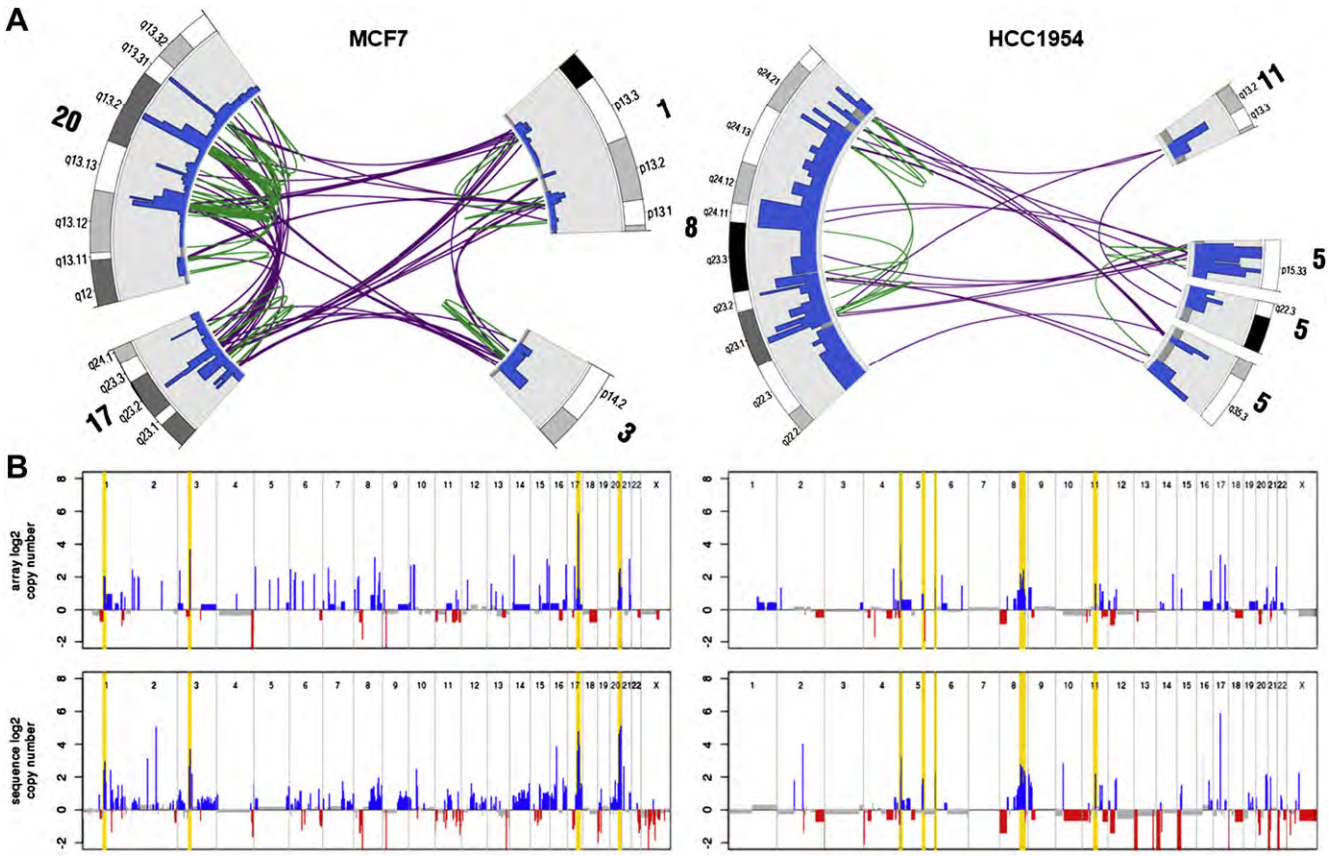

Figure 2 (A) Arc visualizations of the largest MCF7 and HCC1954 breakpoint cliques and their association with copy number amplification. Chromosome cytobands are shaded and labeled. The colored rearrangements depict breast tumor intrachromosomal (green) and interchromosomal (purple) mutations. Copy number counts are plotted with gains in blue, losses in red, and normal diploid in grey (count scales are from zero to MCF7: max = 60; HCC1954: max = 15). (B) MCF7 and HCC1954 $\log_2$ copy number plots of Affymetrix 100K SNP chip arrays (34) (top) and Illumina mate pair mapped sequence counts (bottom); gains are plotted in blue, losses in red, and normal diploid in grey. Highlighted regions correspond to the largest breakpoint cliques from A.

DSBR genes may have contributed to different patterns of genomic instability. For example, in MCF7 we identified the gene–gene fusion of *RAD51C* exons 1–7 to the neuronal-specific gene *ATXN7* exons 6–13 resulting in an expressed chimeric transcript. *RAD51C* is a paralog of *RAD51* a gene central to DNA DSBR. *RAD51C* is an essential component of a complex reported to be involved in resolving HJs (29) formed during DSBR (38) and, as such, is integral to the maintenance of genomic stability. The translocation we have identified eliminates the domain of *RAD51C* that binds other family members such as *RAD51D* and *XRCC3* (39), possibly disrupting formation of the complex responsible for resolving HJs.

Also in MCF7, we identified a truncation of the *BRIP1* gene, *BRCA1*-interacting protein-1. *BRIP1* was originally identified as a helicase-like protein that interacts directly with *BRCA1* and contributes to its DNA repair function. *BRIP1* binds to the BRCT repeat in *BRCA1.* The C-terminus of *BRIP1* is critical for its interaction with *BRCA1,* and a truncation mutant has been shown to block DSBR (40–42). Clinically, germline truncation mutations of *BRIP1* have been identified in familial breast cancer without mutations of *BRCA1/2,* and *BRIP1* truncations confer a twofold increased risk of developing breast cancer. We identified a translocation that results in the loss of the last three exons (exons 18–20); however, the fused DNA (3p14) downstream of *BRIP1* does not contain any exons or introns. The truncation at exon 17 of *BRIP1* would eliminate the C-terminal third of *BRIP1* and eliminate binding to *BRCA1.* However, it is unclear at present whether the truncated mRNA would be stable because there is no transcription stop site or polyA tail.

In HCC1954 we discovered a novel gene truncation of *UIMC1* (also referred to as the *BRCA1-A* complex subunit *RAP80*). *RAP80* has been extensively studied because of its roles in localizing *BRCA1* to DNA double strand break sites, regulating *BRCA1*-dependent DNA damage checkpoint function, and as a potential tumor suppressor (43,44). Whereas many *RAP80* missense SNP mutations have been identified in non-*BRCA1/2* multiethnic breast cancer cases (45,46), no truncating mutation of the *RAP80* gene in breast cancer has been previously published. Interestingly, Dr. Xiaochun Yu has identified a truncating SNP mutation on *RAP80* cDNA (G1107A) in the ovarian adenocarcinoma cell line TOV21G that results in a premature stop codon at Trp369. This truncation product disrupts the *RAP80* interaction with *BRCA1* and fails to localize to nuclear foci after DNA damage (47). The *UIMC1* truncation we identified cleaves the native transcript after exon 10 and results in loss

of the C-terminus exons 11−15, similarly eliminating DNA recognition and binding capability.

Although our fosmid diTags and Illumina mate pairs do not detect the previously published HCC1954 gene truncation of *MRE11A* (33), we did confirm the existence of the t(4;11)(q32;q21) genomic lesion involving *MRE11A* in another study by using 2 kb Life Technologies SOLiD mate pairs (unpublished data). *MRE11A* is involved in homologous recombination, telomere length maintenance and DNA DSBR; and this truncation eliminates its DNA binding domain in the HCC1954 breast cancer cell line.

Coinciding occurrences of rearrangement clustering and amplification point to similar histories of genomic instability in MCF7 and HCC1954

As evidenced from Figure 1A, the breakpoints in MCF7 and HCC1954 are not evenly distributed across the genome. A number of clusters of closely spaced breakpoints are evident. To formally delineate clustered breakpoints from the remainder, breakpoints within 2 Mb in MCF7 and 5 Mb in HCC1954 were clustered. In each cell line, the cluster containing the highest number of breakpoints was selected to seed a connected graph where chromosome segments are nodes and spanning breakpoints are edges. In MCF7, four clusters emerged at cytobands 1p13.1−p21.1, 3p14.1−p14.2, 17q22−q24.3, and 20q12−q13.33. In HCC1954, five clusters emerged at cytobands 5p15.3, 5q22.3−q23.2, 5q35.2−q35.3, 8q22.2−q24.22, and 11q13.2−q12.3. Moreover, the four MCF7 and five HCC1954 clustered breakpoint locations coincide exactly with high-level amplifications in their respective genomes, indicating possible positive selection and functional significance (Figure 2).

The amplification patterns found in MCF7 and HCC1954 are consistent with the complex firestorm pattern described by Hicks et al. that associate with breast cancer prognostic markers and correspond with poor patient outcomes (48). Interestingly, the most often detected recurrent locations of firestorm amplification identified by Hicks et al. within the 243 breast tumors studied, reside on chromosomal arms 11q and 17q. These loci contain the genes *CCND1* on 11q and *ERBB2* on 17q, noted previously to be highly amplified in HCC1954, and may drive selection for these amplifying mutations.

In both cell lines, the remaining nonclustered or dispersed breakpoints were highly associated with LCRs. The dispersed breakpoints in MCF7 show a 9.8-fold enrichment for LCRs, while the trend is reiterated in HCC1954 with a 9.1-fold enrichment for LCRs. LCR enrichment at dispersed breakpoints is a characteristic previously described in MCF7 (28), and is recurrently identified in HCC1954. This finding is in contrast to the clustered breakpoints, which do not exhibit enrichment for LCRs.

Discussion

It is known that chromosomal rearrangements are highly associated with repetitive sequences in genomic disorders and cancer. Up to a quarter of entries in the Gross Rearrangement Breakpoint Database (http://www.uwcm.ac.uk/uwcm/mg/grabd) show presence of repetitive elements (49). The repetitive elements range in size and may be as large as 6 kb in the case of long interspersed nuclear elements and may cluster, creating long stretches of nonunique sequence. Breakpoints that overlap repetitive sequence elements may not be detected by 5 kb (or shorter-range) mate pair libraries. Even if the breakpoint is detected, the nonunique sequence surrounding the rearrangement may make validation by PCR challenging. Having large clonal-sized inserts, such as fosmid diTags overcome this problem by spanning repetitive sequences and correctly identifying aberrant fusions. For example, in our previous study of MCF7 cells, we identified the expressed *DEPDC1B−ELOVL2* chimeric mRNA transcript, which is formed by a 5q12.1 intrachromosomal inversion (28). This breakpoint was detected by using fosmid diTags, but not 5 kb sized mate pairs as a result of presence of long or short interspersed nuclear elements and microsatellites surrounding the site of rearrangement.

In many cases, optimal PCR primer design is hindered by the presence of repetitive sequence surrounding the join. This is common when rearrangements are facilitated by homologous recombination (50,51). Short repetitive elements or longer segmental duplications (also referred to as LCRs) at sites of rearrangement severely limit the number of unique priming positions. Fosmid-sized inserts are able to span such repetitive regions, thus providing a means of validating breakpoints even in cases of PCR assay failure. For example, there are two previously published gene truncations identified by our fosmid diTag and 5 kb mate pair libraries that failed breakpoint spanning PCR assay confirmation. First is the t(5;8)(q35.3;q24.21) translocation in HCC1954 involving the truncation of *NSD1,* a fusion protein also found in myeloid leukemia (33). Second is the t(3;15)(p14.1;q23.2) translocation in MCF7 involving the truncation of *BRIP1,* a *BRCA1*-interacting protein that contributes to DNA repair (28). Although these two gene truncations were cross-validated by fosmid and 5 kb sized inserts, PCR assay across the breakpoint resulted in amplification failure. In these cases, breakpoint spanning primer design was hindered as a result of the presence of interspersed nuclear elements and long terminal repeats across the aberrant joins.

We showed that fosmid-sized inserts are adept at spanning repetitive sequences known to exist at sites of gross rearrangement and LCRs associated with homologous recombination. Combining fosmid diTag and 5 kb Illumina mate pair libraries we were able to detect and validate aberrant fusions involving repetitive genomic sequence where detection by shorter end sequence profiles alone or validation by breakpoint spanning PCR assays failed. In addition, we observed that those rearrangements detected by both insert size ranges exhibit threefold enrichment for cancer-specific somatic mutation and twofold reduction in false-positive detection when compared to the 5 kb mate pairs alone.

For those breast cancer-specific somatic mutations involving genes, we queried transcriptome fusion and truncation literature to corroborate our finding and assess the extent to which our combined fosmid diTag and 5 kb mate pair libraries rediscovered known chimeric transcripts reported in MCF7 and HCC1954. We identified genomic

alterations corresponding to upward of approximately half of the published MCF7 and HCC1954 chimeric mRNA transcripts, but it is difficult to assess the lower bound of our sensitivity because it is unclear whether the undetected transcript mutations are due to transsplicing or similar post-transcriptional modifications.

We integrated read density and breakpoint information from mapped fosmid diTags and 5 kb mate pairs to accurately identify distinct copy number variation in MCF7 and HCC1954. We discovered distinct driver oncogenes associated with high copy number amplifications in MCF7 and HCC1954. The distinct structural mutability profiles between MCF7 and HCC1954 correlate to their phenotypic differences. Amplified chromosomal segments, breakpoint clusters, and affected genes are located at different positions across the MCF7 and HCC1954 genomes; and correspond to overexpression of different oncogenes, silencing of diverse tumor suppressors, and distinct defects in DNA repair machinery responsible for homology-driven repair of double-stranded DNA breaks. It is intriguing that in conjunction with mutations in the same DNA repair pathway we also find similar patterns of structural mutability in the two cell lines. Both have clustered and dispersed breakpoints; both exhibit clustered breakpoints in regions of high copy number amplification and dispersed breakpoints that are enriched for the presence of LCRs.

Acknowledgments

This project was funded by the NIH-NHGRI grant 1 R01 HG02583 and NIH-NCI grants R33 CA114151 and R21 CA128496 to AM.

Supplementary data

Supplementary data associated with this article can be found in the online version at 10.1016/j.cancergen.2011.07.009.

References

1. Bignell GR, Santarius T, Pole JC, et al. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. Genome Res 2007;17:1296–1303.
2. Kidd JM, Cooper GM, Donahue WF, et al. Mapping and sequencing of structural variation from eight human genomes. Nature 2008;453:56–64.
3. Tuzun E, Sharp AJ, Bailey JA, et al. Fine-scale structural variation of the human genome. Nat Genet 2005;37:727–732.
4. Volik S, Raphael BJ, Huang G, et al. Decoding the fine-scale structure of a breast cancer genome and transcriptome. Genome Res 2006;16:394–404.
5. Volik S, Zhao S, Chin K, et al. End-sequence profiling: sequence-based analysis of aberrant genomes. Proc Natl Acad Sci U S A 2003;100:7696–7701.
6. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53–59.
7. Campbell PJ, Stephens PJ, Pleasance ED, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 2008;40:722–729.
8. Korbel JO, Urban AE, Affourtit JP, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science 2007;318:420–426.
9. Maher CA, Palanisamy N, Brenner JC, et al. Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci U S A 2009;106:12353–12358.
10. McKernan KJ, Peckham HE, Costa GL, et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 2009;19:1527–1541.
11. Pleasance ED, Cheetham RK, Stephens PJ, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 2010;463:191–196.
12. Stephens PJ, McBride DJ, Lin ML, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature 2009;462:1005–1010.
13. Berger MF, Lawrence MS, Demichelis F, et al. The genomic complexity of primary human prostate cancer. Nature 2011;470:214–220.
14. Hillmer AM, Yao F, Inaki K, et al. Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. Genome Res 2011;21:665–675.
15. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods 2011;8:61–65.
16. Teague B, Waterman MS, Goldstein S, et al. High-resolution human genome structure by single-molecule analysis. Proc Natl Acad Sci U S A 2010;107:10848–10853.
17. Das SK, Austin MD, Akana MC, et al. Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. Nucleic Acids Res 2010;38:e177.
18. Gnerre S, Maccallum I, Przybylski D, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 2011;108:1513–1518.
19. Fullwood MJ, Wei CL, Liu ET, et al. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res 2009;19:521–532.
20. Ruan Y, Ooi HS, Choo SW, et al. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). Genome Res 2007;17:828–838.
21. Chen J, Kim YC, Jung YC, et al. Scanning the human genome at kilobase resolution. Genome Res 2008;18:751–762.
22. Raghavendra NK, Rao DN. Exogenous AdoMet and its analogue sinefungin differentially influence DNA cleavage by R.EcoP15I—usefulness in SAGE. Biochem Biophys Res Commun 2005;334:803–811.
23. Korbel JO, Abyzov A, Mu XJ, et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol 2009;10:R23.
24. Chen K, Wallis JW, McLellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 2009;6:677–681.
25. Ye K, Schulz MH, Long Q, et al. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 2009;25:2865–2871.
26. Miller CA, Hampton O, Coarfa C, et al. ReadDepth: a Parallel R package for detecting copy number alterations from short sequencing reads. PLoS One 2011;6:e16327.
27. Bailey JA, Yavor AM, Massa HF, et al. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res 2001;11:1005–1017.
28. Hampton OA, Den Hollander P, Miller CA, et al. A sequence-level map of chromosomal breakpoints in the MCF-7 breast

cancer cell line yields insights into the evolution of a cancer genome. Genome Res 2009;19:167–177.
29. Mootha VK, Lepage P, Miller K, et al. Identification of a gene-causing human cytochrome c oxidase deficiency by integrative genomics. Proc Natl Acad Sci U S A 2003;100:605–610.
30. Inoue K, Lupski JR. Molecular mechanisms for genomic disorders. Annu Rev Genomics Hum Genet 2002;3:199–242.
31. Maher CA, Kumar-Sinha C, Cao X, et al. Transcriptome sequencing to detect gene fusions in cancer. Nature 2009;458: 97–101.
32. Inaki K, Hillmer AM, Ukil L, et al. Transcriptional consequences of genomic structural aberrations in breast cancer. Genome Res 2011;21:676–687.
33. Zhao Q, Caballero OL, Levy S, et al. Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. Proc Natl Acad Sci U S A 2009;106:1886–1891.
34. Hu X, Stern HM, Ge L, et al. Genetic alterations and oncogenic pathways associated with breast cancer subtypes. Mol Cancer Res 2009;7:511–522.
35. Collins C, Volik S, Kowbel D, et al. Comprehensive genome sequence analysis of a breast cancer amplicon. Genome Res 2001;11:1034–1042.
36. Huang G, Krig S, Kowbel D, et al. *ZNF217* suppresses cell death associated with chemotherapy and telomere dysfunction. Hum Mol Genet 2005;14:3219–3225.
37. Bieche I, Olivi M, Nogues C, et al. Prognostic value of *CCND1* gene status in sporadic breast tumours, as determined by real-time quantitative PCR assays. Br J Cancer 2002;86:580–586.
38. Liu Y, Tarsounas M, O'Regan P, et al. Role of *RAD51C* and *XRCC3* in genetic recombination and DNA repair. J Biol Chem 2007;282:1973–1979.
39. Miller KA, Sawicka D, Barsky D, et al. Domain mapping of the Rad51 paralog protein complexes. Nucleic Acids Res 2004;32: 169–178.
40. Lewis AG, Flanagan J, Marsh A, et al. Mutation analysis of *FANCD2, BRIP1/BACH1, LMO4* and *SFN* in familial breast cancer. Breast Cancer Res 2005;7:R1005–R1016.
41. Yu X, Chini CC, He M, et al. The BRCT domain is a phospho-protein binding domain. Science 2003;302:639–642.
42. Cantor SB, Bell DW, Ganesan S, et al. BACH1, a novel helicase-like protein, interacts directly with *BRCA1* and contributes to its DNA repair function. Cell 2001;105:149–160.
43. Kim H, Chen J, Yu X. Ubiquitin-binding protein RAP80 mediates *BRCA1*-dependent DNA damage response. Science 2007;316: 1202–1205.
44. Liu Z, Wu J, Yu X. CCDC98 targets *BRCA1* to DNA damage sites. Nat Struct Mol Biol 2007;14:716–720.
45. Akbari MR, Ghadirian P, Robidoux A, et al. Germline *RAP80* mutations and susceptibility to breast cancer. Breast Cancer Res Treat 2009;113:377–381.
46. Novak DJ, Sabbaghian N, Maillet P, et al. Analysis of the genes coding for the *BRCA1*-interacting proteins, *RAP80* and Abraxas (*CCDC98*), in high-risk, non-*BRCA1/2*, multiethnic breast cancer cases. Breast Cancer Res Treat 2009;117:453–459.
47. Yu X. Characterize *RAP80*, a potential tumor suppressor gene. University of Michigan - Ann Arbor, U.S. Army Medical Research and Materiel Command; 2009. p. 24.
48. Hicks J, Krasnitz A, Lakshmi B, et al. Novel patterns of genome rearrangement and their association with survival in breast cancer. Genome Res 2006;16:1465–1479.
49. Abeysinghe SS, Chuzhanova N, Krawczak M, et al. Translocation and gross deletion breakpoints in human inherited disease and cancer I: nucleotide composition and recombination-associated motifs. Hum Mutat 2003;22:229–244.
50. Bashir A, Liu YT, Raphael BJ, et al. Optimization of primer design for the detection of variable genomic lesions in cancer. Bioinformatics 2007;23:2807–2815.
51. Bashir A, Lu Q, Carson D, et al. Optimizing PCR assays for DNA-based cancer diagnostics. J Comput Biol;17:369–381.
52. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. Genome Res 2009;19: 1639–1645.
53. Barlund M, Monni O, Weaver JD, et al. Cloning of *BCAS3* (17q23) and *BCAS4* (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. Genes Chromosomes Cancer 2002;35:311–317.
54. Huang GL, Zhang XH, Guo GL, et al. Clinical significance of miR-21 expression in breast cancer: SYBR-Green I–based real-time RT-PCR study of invasive ductal carcinoma. Oncol Rep 2009;21:673–679.
55. Heinonen H, Nieminen A, Saarela M, et al. Deciphering downstream gene targets of PI3K/mTOR/p70S6K pathway in breast cancer. BMC Genomics 2008;9:348.
56. Tamura K, Makino A, Hullin-Matsuda F, et al. Novel lipogenic enzyme ELOVL7 is involved in prostate cancer growth through saturated long-chain fatty acid metabolism. Cancer Res 2009; 69:8133–8140.
57. Liu L, Gong G, Liu Y, et al. Multi-species comparative mapping in silico using the COMPASS strategy. Bioinformatics 2004;20: 148–154.